

Spike-and-Slab Priors for Function Selection in Structured Additive Regression Models

Fabian Scheipl, Ludwig Fahrmeir, Thomas Kneib *

Structured additive regression provides a general framework for complex Gaussian and non-Gaussian regression models, with predictors comprising arbitrary combinations of nonlinear functions and surfaces, spatial effects, varying coefficients, random effects and further regression terms. The large flexibility of structured additive regression makes function selection a challenging and important task, aiming at (1) selecting the relevant covariates, (2) choosing an appropriate and parsimonious representation of the impact of covariates on the predictor and (3) determining the required interactions. We propose a spike-and-slab prior structure for function selection that allows to include or exclude single coefficients as well as blocks of coefficients representing specific model terms. A novel multiplicative parameter expansion is required to obtain good mixing and convergence properties in a Markov chain Monte Carlo simulation approach and is shown to induce desirable shrinkage properties. In simulation studies and with (real) benchmark classification data, we investigate sensitivity to hyperparameter settings and compare performance to competitors. The flexibility and applicability of our approach are demonstrated in an additive piecewise exponential model with time-varying effects for right-censored survival times of intensive care patients with sepsis. Geoadditive and additive mixed logit model applications are discussed in an extensive appendix.

Key-words: parameter expansion, penalized splines, stochastic search variable selection, generalized additive mixed models, spatial regression

1. INTRODUCTION

Recent research on function selection mostly considers the additive model $y = f_1(x_1) + \dots + f_q(x_q) + \epsilon$ for Gaussian responses, sometimes including additional lin-

*Fabian Scheipl is Postdoctoral Fellow (E-mail: fabian.scheipl@stat.uni-muenchen.de) and Ludwig Fahrmeir is Professor (emeritus), Department of Statistics, Ludwig-Maximilians-Universität München, Munich, Germany. Thomas Kneib is Professor, Department of Economics, Georg-August-Universität Göttingen, Göttingen, Germany. This work was supported by the German Science Foundation (DFG grant FA 128/5-1).

ear effects or interactions of functions. In this paper we introduce a spike-and-slab prior structure to perform Bayesian inference and function selection in structured additive regression (STAR) models, i.e., in exponential family regression models with additive predictors incorporating different types of functions or effects. Compared to additive models, we therefore extend function selection to regression with non-Gaussian, in particular discrete responses and the predictor may contain additional components, such as varying coefficient terms $uf(x)$, smooth interactions $f(x_1, x_2)$, spatial effects $f_{\text{geo}}(s)$ for geoadditive regression, and cluster-specific random effects. Functions of continuous covariates are represented through penalized (tensor product) splines, $f_{\text{geo}}(s)$ through (conditionally) Gaussian Markov random fields, and cluster-specific effects through (conditionally) Gaussian i.i.d. random effects, but other basic function expansions, surface smoothers, and spatial models are possible as well, see e.g., Fahrmeir et al. (2004) for details. Any generalized structured additive regression model can then be written in unifying form as

$$E(\mathbf{y}|\boldsymbol{\eta}) = h(\boldsymbol{\eta}), \quad \boldsymbol{\eta} = \mathbf{f}_1 + \dots + \mathbf{f}_P = \mathbf{Z}_1\boldsymbol{\delta}_1 + \dots + \mathbf{Z}_P\boldsymbol{\delta}_P, \quad (1)$$

where the conditional expectation $E(\mathbf{y}|\boldsymbol{\eta})$ of the response vector $\mathbf{y} = (y_1, \dots, y_n)'$ is related to a predictor vector $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)'$ via a known response function h as in generalized linear models. The predictor vector $\boldsymbol{\eta}$ is additively composed of covariate effects $\mathbf{f}_j = (f_j(x_{j1}), \dots, f_j(x_{jn}))'$ of different types. Each effect \mathbf{f}_j can also be a function of multiple covariates and is represented by suitable design matrices \mathbf{Z}_j of dimension $(n \times D_j)$, and D_j -dimensional regression coefficients $\boldsymbol{\delta}_j | s_j^2 \sim N(\mathbf{0}, s_j^2 \mathbf{P}_j^-)$ with fixed positive (semi-)definite scaled precision matrix \mathbf{P}_j and prior variance s_j^2 . Singular precision matrices result from Bayesian P-Splines (Brezger and Lang, 2006) or intrinsic Markov random fields (Rue and Held, 2005).

Section 2.1 shows how the terms in (1) can be reparameterized in terms of modified design matrices associated with conditionally Gaussian i.i.d. coefficient vectors. This reparameterization essentially follows ideas similar to those considered in Ruppert et al. (2003) or Fahrmeir et al. (2004) but is especially designed to improve the selection proper-

ties of our approach. It also has the advantage that additional function selection questions can be handled, for example differentiating between no effect, linear effect and inherently nonlinear effect of a continuous covariate (see Section 2.1 for details).

In function selection, we are interested in finding simple special cases of (1), where some of the functions are identified as having negligible impact on the response. For example, in our geoadditive regression model of rents in the city of Munich, we want to select a subset from a large set of categorical covariates and to decide if effects of age of a building and floor space of the flat are nonlinear or linear, if an interaction between them is necessary, and if a spatial effect in form of a Markov random field representing the location of buildings is needed. In the application dealing with the analysis of survival times of patients that acquired a septic infection after surgery, we are interested in finding a parsimonious model to indicate which covariates have (potentially nonlinear) impact on survival and to evaluate the presence or absence of time-varying effects indicating non-proportional hazards.

In a Bayesian or mixed models framework, functional effects in structured additive regression are reparameterized as i.i.d. Gaussian random effects (or i.i.d. Gaussian priors from a Bayesian perspective). This idea has become quite popular since it allows treating complex regression models as mixed models and makes corresponding restricted maximum likelihood (REML) estimation available for the determination of smoothing variances. Wand (2000) was among the first to recognize this possibility and introduced mixed model based inference for penalized spline regression based on truncated power series expansions in Gaussian regression models. Ruppert et al. (2003) provide an in-depth overview on mixed model based semiparametric regression and describe extensions to the estimation of interaction surfaces and spatial effects. In an (empirical) Bayes framework, the mixed model representation corresponds to a reparameterisation of the prior and REML estimation is interpreted as marginal likelihood estimation, see Fahrmeir et al. (2004) for more details in the context of (generalized) STAR models and Crainiceanu et al. (2005) for the implementation of such models via WinBUGS. The mixed model perspective on semiparametric regression has also led to the development of likelihood ratio

tests that implement the selection of functions $f_j(x_j)$ based on testing $H_0 : s_j^2 = 0$ versus $H_A : s_j^2 > 0$ (cf. Crainiceanu et al., 2005; Greven et al., 2008; Scheipl et al., 2008), since $s_j^2 = 0$ implies $f_j = \mathbf{Z}_j \delta_j = \mathbf{0}$. However, these likelihood ratio tests are so far only applicable in Gaussian regression models and are not suitable for automatic function selection in complex regression models with a large number of potentially nonlinear effects.

To develop a Bayesian counterpart of likelihood ratio tests, it seems natural to impose a bimodal spike-and-slab prior on the variances s_j^2 , as suggested in Ishwaran and Rao (2005) for the case of variable selection in high-dimensional linear models, i.e. for selecting single scalar regression coefficients. Indeed, our first attempt to select functions in STAR models was based on this simple idea. However, as shown in the web appendix, such a straightforward approach is rendered infeasible by the severe convergence and mixing problems it causes. Informally, the problem is that a small variance for a block of coefficients implies small coefficient values and small coefficient values in turn imply a small variance. Therefore, blockwise MCMC samplers are unlikely to exit a basin of attraction around zero. We therefore propose a multiplicative parameter expansion for the regression coefficients inspired by the work of Gelman et al. (2008) in the context of mixed models. We show that this parameter expansion leads to an efficient MCMC strategy and to a prior with regularization properties similar to \mathcal{L}_q -penalization with $q < 1$ in Section 2.3.

The main advantages of our new prior structure can be summarized as follows: First, unlike standard stochastic search variable selection approaches, it can routinely be used with non-Gaussian responses from the exponential family based on iteratively weighted least squares updates. Second, it supports the full generality of STAR models, i.e. it accommodates all types of regularized effects with a (conditionally) Gaussian prior such as simple covariates or covariate blocks (both continuous and categorical), penalized splines (uni- or multivariate), spatial effects, random effects or ridge-penalized factors and all their interactions (e.g. (space-)varying coefficient terms or random slopes). Third, it scales to data sets of intermediate size with thousands of observations and high-dimensional predictors including hundreds of model terms. Fourth, it is implemented

in publicly available and user-friendly open source software (R-package `spikeSlabGAM` (Scheipl, 2011b)), therefore allowing reproducibility of our results and immediate application to new data sets.

Due to the practical importance of the topic, there is a vast amount of literature on selecting components in predictors of regression models. Most previous work considers selection of variables or associated (single) regression coefficients in high-dimensional (generalized) linear models. Penalization approaches have become quite popular, in particular the Lasso or the SCAD penalty and modifications as the adaptive or group Lasso. Another branch is boosting, see Bühlmann and Hothorn (2007) for a survey. Most Bayesian approaches for variable selection are based on spike-and-slab priors for regression coefficients, see for example the stochastic search variable selection (SSVS) approach in George and McCulloch (1993), among other methods, and the review in O’Hara and Sillanpää (2009).

In comparison, research on function selection is more sparse. Recently, Marra and Wood (2011) have proposed a “double shrinkage” approach for GAMs with an additional penalty on the null space of the smoothness penalty which enables shrinking entire functional terms to zero. Most other penalization methods only consider additive models for continuous (Gaussian) responses and perform function selection by penalizing certain norms of functional components or associated blocks of basis function coefficients in Lasso- or SCAD-type fashion. Lin and Zhang (2006) proposed the component selection and smoothing operator (COSSO) in additive smoothing spline ANOVA models. Motivated by the adaptive group Lasso, Storlie et al. (2010) propose the adaptive COSSO (ACOSSO) to penalize each functional component differently so that more flexibility is obtained. Its superior performance to the COSSO (and MARS) is demonstrated for simulated and real data. A similar adaptive group Lasso approach is studied in Huang et al. (2010). Ravikumar et al. (2009) estimate sparse additive models by penalizing the quadratic norm of functional predictor terms, Meier et al. (2009) additionally impose a smoothness penalty.

Many Bayesian function selection approaches are based on introducing spike-and-slab

priors with a point mass at zero directly for blocks of basis function coefficients or, equivalently, indicator variables for functions being zero or nonzero. Wood et al. (2002) and Yau et al. (2003) describe implementations using a data-based prior that requires two MCMC runs, a pilot run to obtain a data-based prior for the “slab” part and a second one to estimate parameters and select model components. Panagiotelis and Smith (2008) combine this stochastic search variable selection approach with partially improper Gaussian priors, as for basis coefficients of Bayesian P-splines, in high-dimensional additive models. They suggest several sampling schemes that dominate the scheme in Yau et al. (2003). A more general approach based on double exponential regression models that also allows for flexible modeling of the dispersion is described by Cottet et al. (2008). They use a reduced rank representation of cubic smoothing splines with a very small number of basis functions to model the smooth terms in order to reduce the complexity of the fitted models, and, presumably, to avoid the mixing problems already mentioned and described in the web appendix. Reich et al. (2009) use the smoothing spline ANOVA framework and a spike-and-slab prior for the variance of Gaussian process priors to perform variable and function selection via SSVS for Gaussian responses. In a wavelet-based functional Gaussian mixed model framework, Morris and Carroll (2006) place spike and slab priors directly on wavelet coefficients to decide whether they are important for representing functional effects. This approach is further extended by Zhu et al. (2011) who improve robustness and adaptivity by considering scale mixtures of normals in the spike and slab specification. Zhu et al. (2010) develop a probit model for functional data classification that allows to select functional predictors. Their hierarchical model is based on a latent Gaussian model and on Gaussian process priors for functional effects. Like Reich et al. (2009), they place a spike and slab prior on the variance, while the remaining part of the Gaussian process covariance matrix is assumed to be known. This (strong) assumption and the possibility to integrate out functional effect parameters in full conditionals for variance parameters facilitates MCMC inference in this case.

Function selection in generalized additive models using Bayes factors has recently been proposed in two ways: Sabanés Bové and Held (2011) present a basis function selec-

tion approach for Bayesian fractional polynomials for potentially nonlinear effects, while the approach of Sabanés Bové et al. (2011) is based on a grid of fixed effective degrees of freedom for each penalised spline. Chipman et al. (2010) propose Bayesian adaptive regression trees (BART) to develop a completely non-parametric prediction-oriented approach.

In principle, adaptive smoothing approaches based on knot selection strategies as suggested for example in Denison et al. (1998) for the Bayesian MARS or Dimatteo et al. (2001) for adaptive regression splines (BARS) could be considered as additional competitors for function selection. In particular, knot selection strategies typically involve the possibility to deselect covariate effects by excluding all basis functions associated with a specific covariate. However, this possibility is usually only a by-product of the model specification and is not directly intended for function selection. Roughly, these approaches correspond to specifying separate i. i. d. spike and slab priors for the scalar basis function coefficients, rather than imposing a multivariate spike and slab prior on the entire vector of basis function coefficients as in Panagiotelis and Smith (2008). Note that even fitting of a single function with adaptive smoothing based on knot selection can be extremely time-consuming, see for example the comparison of BARS with the adaptive penalty approach in Krivobokova et al. (2008). As a consequence, selecting functions in this fashion becomes inefficient, if not computationally infeasible. Therefore we will not consider knot-selection approaches in the rest of this paper.

2. NMIG PRIORS FOR FUNCTION SELECTION

2.1. Generic Parameterization

Many of the regression terms available for STAR models are associated with conditionally Gaussian priors with zero mean and general positive semidefinite precision matrices \mathbf{P} (cf. (1)). For example, in the case of penalized splines the precision matrix represents the dependence structure imposed by the random walk prior (Brezger and Lang, 2006) or in the case of Gaussian Markov random fields, the precision matrix is defined by the

neighborhood structure underlying the geographical arrangement of the data (Rue and Held, 2005). We will now show that such general Gaussian priors can always be recast based on i.i.d. priors:

Assume that $\delta|s^2 \sim N(\mathbf{0}, s^2 \mathbf{P}^-)$ represents the D -dimensional regression coefficient corresponding to one of the terms $f(\mathbf{z}) = \mathbf{Z}\delta$ appearing in a STAR model (1). Let K denote the dimension of the null-space of \mathbf{P} . Since $f(\mathbf{z}) = \mathbf{Z}\delta$ and $\mathbf{Z}\delta|s^2 \sim N(\mathbf{0}, s^2 \mathbf{Z}\mathbf{P}^-\mathbf{Z}')$, the spectral decomposition $\mathbf{Z}\mathbf{P}^-\mathbf{Z}' = \mathbf{U}\mathbf{V}\mathbf{U}'$ with orthonormal \mathbf{U} and a diagonal \mathbf{V} with entries ≥ 0 yields an orthogonal basis representation for the improper prior covariance of $f(\mathbf{z})$. For \mathbf{Z} with D columns and full column rank and \mathbf{P} with rank $d = D - K$, all eigenvalues in \mathbf{V} except the first d are zero. Now write $\mathbf{Z}\mathbf{P}^-\mathbf{Z}' = [\mathbf{U}_+ \mathbf{U}_0]' \begin{bmatrix} \mathbf{V}_+ & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} [\mathbf{U}_+ \mathbf{U}_0]$, where \mathbf{U}_+ is a matrix of eigenvectors associated with the positive eigenvalues in \mathbf{V}_+ , and \mathbf{U}_0 are the eigenvectors associated with the zero eigenvalues. With $\mathbf{X}_{\text{pen}} = \mathbf{U}_+ \mathbf{V}_+^{1/2}$ and $\boldsymbol{\beta}_{\text{pen}} \sim N(\mathbf{0}, v^2 \mathbf{I})$, $f_{\text{pen}}(\mathbf{z}) = \mathbf{X}_{\text{pen}} \boldsymbol{\beta}_{\text{pen}}$ has a proper Gaussian distribution that is proportional to that of the partially improper prior of $f(\mathbf{z})$ (Rue and Held, 2005, eq. (3.16)) but parameterizes only the penalized component of $f(\mathbf{z})$, while $\mathbf{X}_0 = \mathbf{U}_0$ and the associated coefficients $\boldsymbol{\beta}_0$ parameterize functions $f_0(\mathbf{z}) = \mathbf{X}_0 \boldsymbol{\beta}_0$ in the null space of \mathbf{P} . In summary, we reparameterize and decompose $f(\mathbf{z}) = f_0(\mathbf{z}) + f_{\text{pen}}(\mathbf{z})$. Our reparameterization follows similar ideas as in early references on mixed model based inference in semiparametric regression (see for example Wand (2000); Ruppert et al. (2003) for corresponding frequentist approaches and Fahrmeir et al. (2004); Crainiceanu et al. (2005) for Bayesian interpretations) but is designed for the special purpose of function selection. Basing the decomposition on the spectral decomposition of $\mathbf{Z}\mathbf{P}^-\mathbf{Z}'$ instead of the spectral decomposition of \mathbf{P} yields an orthogonal basis representation and therefore facilitates the differentiation between penalized and unpenalized parts of a function.

In practice, it is unnecessary and impractically slow to compute all n eigenvectors and values for a full spectral decomposition $\mathbf{U}\mathbf{V}\mathbf{U}'$. Only the first d are needed for \mathbf{X} , and of those the first few typically represent most of the variability in $f(\mathbf{z})$. Our implementation makes use of a fast truncated bidiagonalization algorithm (Baglama and Reichel, 2006) available in `irlba` (Lewis, 2009) to compute only the largest d eigenvalues of $\text{Cov}(f(\mathbf{z}))$

and their associated eigenvectors. Only the first d eigenvectors and -values whose sum represents at least .995 of the sum of all eigenvalues are used to construct the reduced rank orthogonal basis \mathbf{X}_{pen} with d columns. E.g. for a cubic P-spline with second order difference penalty and 20 basis functions (i.e. $D = 20$ columns in \mathbf{Z} and $K = 2$), \mathbf{X} will often have only 8 to 12 columns and \mathbf{X}_0 has one column for the linear trend since the constant part of $f_0(z)$ is subsumed into the global intercept to ensure identifiability.

The advantages of applying a reparameterization are two-fold: First, we can separate the penalized part of a predictor function $f(z)$. Second, we can now not only assign an i.i.d. Gaussian prior to the penalized part but also to the unpenalized part. Of course we then no longer have a one-to-one transformation of the original prior but we can perform function selection on both the penalized and the unpenalized parts. For example, in case of penalized splines with k -th order random walk prior, the space of unpenalized functions consists of all polynomials of order less than $k - 1$. Separating these polynomials from the non-polynomial, penalized part of the function opens up the possibility to decide whether a nonlinear effect for a continuous covariate should be included in the model at all, whether it can be described in terms of a linear effect or whether a nonlinear effect is needed. The resulting models are more parsimonious and easier to interpret.

In the following, we assume that the reparameterization has been applied to all relevant model terms and we treat $f_0(z)$ and $f_{\text{pen}}(z)$ as separate model terms, so that $\boldsymbol{\eta} = \boldsymbol{\eta}_0 + \sum_{j=1}^P \mathbf{Z}_j \boldsymbol{\delta}_j$ with $\boldsymbol{\delta}_j | s_j^2 \sim N(\mathbf{0}, s_j^2 \mathbf{P}_j^-)$ (cf. (1)) can now be rewritten as

$$\boldsymbol{\eta} = \boldsymbol{\eta}_0 + \sum_{j=1}^p \mathbf{X}_j \boldsymbol{\beta}_j, \quad (2)$$

with $\boldsymbol{\eta}_0$ containing a global intercept, offset terms, and effects that are not associated with a variable selection prior, $\boldsymbol{\beta}_j | v_j^2 \sim N(\mathbf{0}, v_j^2 \mathbf{I})$ and $p \geq P$ the new number of separate model terms now comprising both $\mathbf{X}_{0,j}$ and $\mathbf{X}_{\text{pen},j}$, $j = 1, \dots, p$.

2.2. Parameter-Expanded NMIG Prior

Inspired by the work of Gelman et al. (2008), we propose a multiplicative parameterization of β_j and combine it with a spike-and-slab prior based on a mixture of inverse gamma distributions for the variances v_j^2 . More specifically, we multiplicatively expand the d_j -dimensional vector β_j to $\beta_j = \alpha_j \xi_j$, $\xi_j \in \mathbb{R}^{d_j}$, where the scalar parameter α_j parameterizes the importance of the j -th coefficient block, while ξ_j “distributes” α_j across the entries in β_j .

We assume that $\alpha_j \sim N(0, v_j^2)$ follows a univariate Gaussian distribution with variance $v_j^2 = \gamma_j \tau_j^2$ given by the product of an indicator variable γ_j and the prior variance τ_j^2 . In a further level of the hierarchy we specify

$$\tau_j^2 \sim \Gamma^{-1}(a_\tau, b_\tau), \quad \gamma_j \sim w\delta_1(\gamma_j) + (1 - w)\delta_{v_0}(\gamma_j),$$

i.e. the variance τ_j^2 is assumed to follow an inverse gamma-prior with shape parameter a_τ and scale parameter b_τ chosen such that $b_\tau \gg a_\tau$. As a consequence, the mode b_τ/a_τ is significantly greater than 1. The indicator γ_j takes the value 1 with probability w or some (very) small value v_0 with probability $1 - w$. The implied prior for the effective variance $v_j^2 = \gamma_j \tau_j^2$ is a bimodal mixture of inverse gamma distributions, with one component strongly concentrated on very small values – the *spike* with $\gamma_j = v_0$ and effective scale parameter $v_0 b_\tau$ – and a second more diffuse component with most mass on larger values – the *slab* with $\gamma_j = 1$ and scale b_τ . A coefficient associated with a variance that is primarily sampled from the *spike*-part of the prior will be strongly shrunk towards zero if v_0 is sufficiently small, so that the posterior probability for $\gamma_j = v_0$ can be interpreted as the probability of exclusion of β_j and the corresponding function f_j from the model. A Beta prior for the mixture weights w can be used to incorporate the analyst’s prior knowledge about the sparsity of β or, more practically, enforce sufficiently sparse solutions for overparameterized models.

We refer to the complete prior structure for α_j as a normal-mixture-of-inverse gamma (NMIG) distributions, denoted as $\alpha_j \sim \text{NMIG}(v_0, w, a_\tau, b_\tau)$. A similar NMIG prior has

originally been suggested in Ishwaran and Rao (2005) for selecting single coefficients $\beta_j \sim N(0, v_j^2)$ in high-dimensional linear models.

Integrating out τ_j^2 and γ_j from $\alpha_j \sim N(0, v_j^2 = \gamma_j \tau_j^2)$, keeping w fixed, the bimodal Inverse Gamma prior induces the mixture of two scaled t-distributions

$$\alpha_j | w \sim (1 - w)t(df, s_0) + w t(df, s_1),$$

with $df = 2a_\tau$, $s_0 = \sqrt{v_0 b_\tau / a_\tau}$, $s_1 = \sqrt{b_\tau / a_\tau}$, as a spike and slab prior directly on α_j . This may suggest to put other spike and slab priors directly on α_j , in particular a bimodal mixture of normals

$$(1 - w)N(0, v_{0j}^2) + wN(0, v_{1j}^2), \quad (3)$$

with $v_{1j}^2 \gg v_{0j}^2$, introduced by George and McCulloch (1993) for selecting scalar coefficients in high dimensional linear models. Another alternative seems to be to replace $N(0, v_{0j}^2)$ by a point mass at zero, but this works only for Gaussian regression models where it is possible to integrate out parameters analytically from full conditionals for indicator variables.

We prefer the NMIG prior for α_j , inducing the student spike and slab prior, for several reasons: First, as noted in Ishwaran and Rao (2005) and supported by our own experience, posterior inference and model selection is relatively robust with respect to hyperparameters in the NMIG hierarchy. It can be more difficult to specify v_{0j}, v_{1j} for the Gaussian spike and slab prior (3). Second, Section 5 of Ishwaran and Rao (2005) provides theoretical arguments in favor of a bimodal continuous prior for variances, as in the NMIG and peNMIG prior, whereas (3) is a bimodal discrete prior $(1 - w)I(v_j^2 = v_{0j}^2) + wI(v_j^2 = v_{1j}^2)$. Furthermore, the favorable shrinkage properties of the NMIG prior induce desirable shrinkage properties for the coefficient vector β_j , see Subsection 2.3.

Entries of the vector ξ_j are assigned the prior distribution

$$\xi_{jk} | m_{jk} \sim N(m_{jk}, 1), \quad m_{jk} \sim \frac{1}{2}\delta_1(m_{jk}) + \frac{1}{2}\delta_{-1}(m_{jk}).$$

i.e. we assume i.i.d. mixtures of two Gaussian distributions with expectation ± 1 . Although the marginal prior for ξ_{jk} still has zero expectation, the bivariate mixture assigns most of the prior mass close to the multiplicative identity (with positive or negative sign). This enables the interpretation of α_j as the “importance” of the j -th coefficient block and yields a marginal prior for β_j that is less concentrated on small absolute values than with a standard assumption like $\xi_{jk} \sim N(0, 1)$.

The prior specification for β_j is completed by assuming prior independence between α_j and ξ_j . We write $\beta_j \sim \text{peNMIG}(v_0, w, a_\tau, b_\tau)$ for the complete prior structure also summarized as a directed acyclic graph in Figure 1.

The main advantage of the peNMIG-prior is that the dimension of the coefficient vector associated with updating γ_j and τ_j^2 is equal to one in every penalization group, since the Markov blankets of both γ_j and τ_j only contain the scalar parameter α_j instead of the vector β_j . This is crucial in order to avoid mixing problems that would arise in a conventional NMIG prior without parameter expansion (see web appendix A). The vector $\xi = (\xi'_1, \dots, \xi'_p)'$ is decomposed into subvectors ξ_j , $j = 1, \dots, p$, associated with the different model terms and their respective entries α_j in α . Note that η_i typically also includes terms that are not under selection, such as known offsets, a global intercept or covariate effects that are forced into the model. Their coefficients are associated with weakly informative flat Gaussian priors.

For Gaussian responses, we assume an $\text{IG}(a_\sigma, b_\sigma)$ prior for the variance σ^2 .

2.3. Shrinkage Properties

This section describes regularization properties of marginal priors for regression coefficients, implied by the hierarchical prior structure described in the previous section and visualized in Figure 1. We analyze marginal priors because it is their shape - and less that of the conditional priors - that determines the shrinkage properties.

For comparison with other shrinkage priors recently suggested for pure variable selection, i.e. for selecting single scalar regression coefficients rather than blocks of coefficients, we first consider *univariate marginal priors*. Omitting indices j and the dependence on hy-

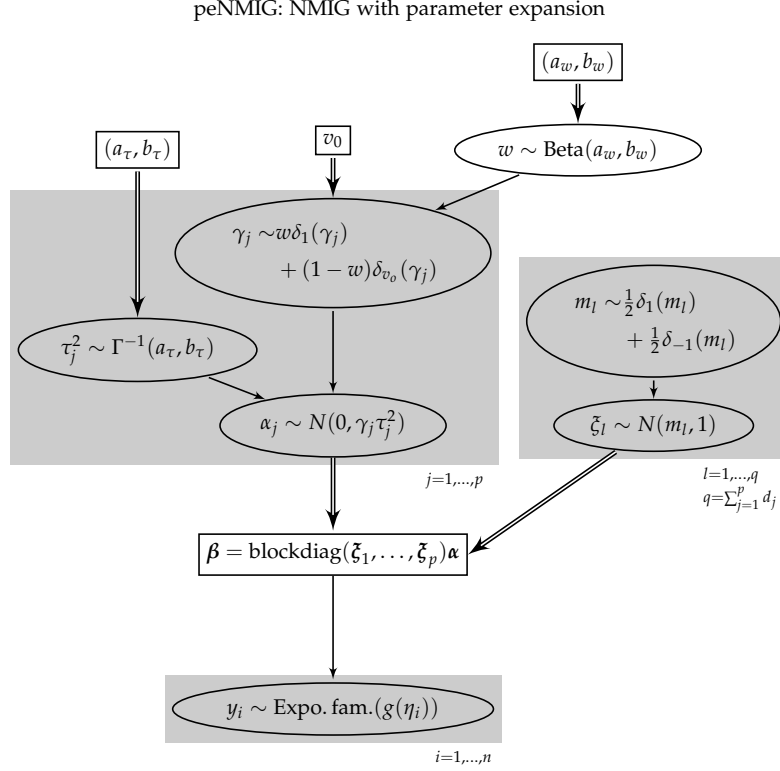


Figure 1: Directed acyclic graph of NMIG model with parameter expansion. Ellipses are stochastic nodes, rectangles are deterministic/logical nodes. Single arrows are stochastic edges, double arrows are logical/deterministic edges.

perparameters a_τ, b_τ, a_w, b_w and v_0 , the marginal prior for a scalar coefficient β is obtained by integrating out all other random variables appearing in the prior hierarchy, i.e.

$$p(\beta = \alpha \xi) = \int p(\alpha | \gamma, \tau^2) p\left(\frac{\beta}{\alpha} | m\right) \frac{1}{|\alpha|} p(m) p(\tau^2) p(\gamma | w) p(w) d\alpha dm d\tau^2 d\gamma dw. \quad (4)$$

Whereas the marginal prior for the original NMIG spike-and-slab prior of Ishwaran and Rao (2005) can be derived analytically as a mixture of scaled t-distributions, the above integral has no known closed form and has to be computed numerically. The marginal NMIG prior has a finite spike around zero, corresponding to the first component of the scaled t-mixture, and a slab corresponding to the second component. In comparison, the marginal peNMIG prior has heavier tails and an infinite spike at zero. Its shape is very close to the horseshoe prior which has favorable theoretical properties. The peNMIG

prior also looks similar to the original spike-and-slab prior suggested by Mitchell and Beauchamp (1988). The tails of the marginal peNMIG prior are also heavy enough to imply redescending score functions which ensures Bayesian robustness of the resulting shrinkage estimators. The shape of the score function is similar to that of an \mathcal{L}_q -prior with $q \rightarrow 0$ and is fairly robust with respect to hyperparameters, see Figure 5 in Scheipl (2010).

In summary, the peNMIG prior for scalar β combines an infinite spike at zero with heavy tails. This desirable combination is similar to other shrinkage priors, including the horseshoe prior and also the normal-Jeffreys prior (Bae and Mallick, 2004), for which both robustness for large values of β and very efficient estimation of sparse coefficient vectors have been shown (Carvalho et al., 2010; Polson and Scott, 2010).

A main advantage of our peNMIG prior is its generalization to *multiple shrinkage* of blocks of coefficients vectors, both in terms of sampling and shrinkage properties. We illustrate this for two-dimensional coefficients (β_1, β_2) , distinguishing two situations: First, β_1 and β_2 come from two different coefficient (or penalization) groups, i.e. we have $\beta_1 = \alpha_1 \zeta_1$ and $\beta_2 = \alpha_2 \zeta_2$, with α_1 independent from α_2 , in terms of the reparameterized coefficients. Second, β_1 and β_2 are from the same block of coefficients representing one penalization group, so that $(\beta_1, \beta_2) = \alpha(\zeta_1, \zeta_2)$, with identical $\alpha_1 = \alpha_2 = \alpha$. Figure 2 shows contour plots of $\log p(\beta_1, \beta_2)$ for these two situations and, for comparison, for the original NMIG prior which applies only to the first situation.

The shape of the constraint region implied by the peNMIG prior (middle of Figure 2) has the convex shape of a \mathcal{L}_q -penalty function with $q < 1$, which has the desirable properties of simultaneous strong shrinkage of small coefficients and weak shrinkage of large coefficients due to its closeness to the \mathcal{L}_0 penalty.

The contours of the NMIG prior (left part of Figure 2) have different shapes depending on the distance from the origin. Close to the origin ($\beta < .3$), they are circular and very closely spaced, implying strong ridge-type shrinkage: Coefficient values close to zero fall into the spike-part of the prior and will be strongly shrunk towards zero. Moving away from the origin ($.3 < \beta < .8$), the shape of the contours defining the constraint region

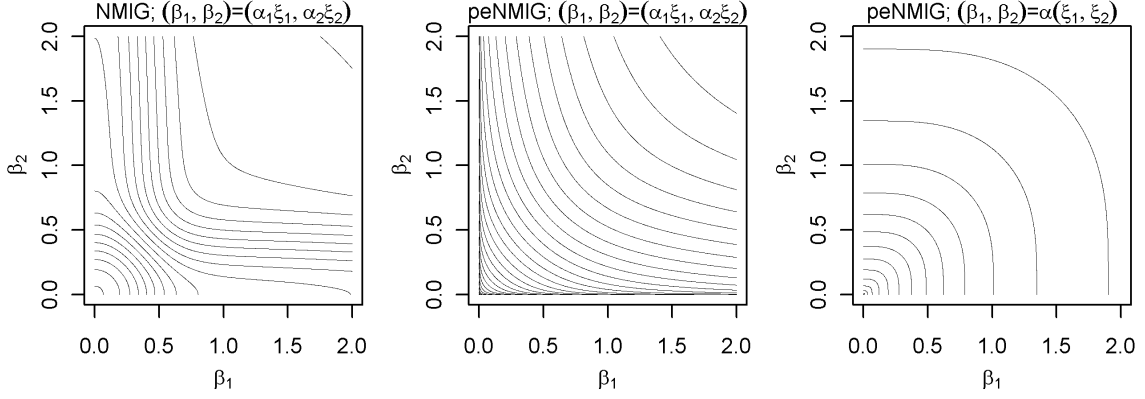


Figure 2: Contour plots of $\log p((\beta_1, \beta_2)')$ for $a_\tau = 5$, $b_\tau = 50$, $v_0 = 0.005$, $a_w = b_w$ for (from left to right) the NMIG prior for two coefficients from different penalization groups, the peNMIG prior for two coefficients from different penalization groups and the peNMIG prior for two coefficients from the same penalization group.

morphs into a rhombus shape with rounded corners that is similar to that produced by a Cauchy prior. Still further from the origin ($1 < \beta < 2$), the contours become convex and resemble those of the contours of an \mathcal{L}_q -penalty function with $q < 1$. Coefficient pairs in this region will be shrunk towards one of the axes, depending on which of their maximum likelihood estimators is bigger and their posterior correlation. For larger values of coefficient pairs, the contours (not shown in Figure 2) imply ridge-type shrinkage.

The shape of the constraint region for coefficient pairs $(\beta_1, \beta_2) = \alpha(\xi_1, \xi_2)$ from the same penalization group (right part of Figure 2) resembles that of a square with rounded corners. Compared with the convex shape of the constraint region in the middle, this shape induces less shrinkage towards the axes and more towards the origin or along the bisecting angle.

2.4. Markov Chain Monte Carlo Algorithm

Posterior inference and function selection is based on a blockwise Metropolis-within-Gibbs sampler. The sampler cyclically updates the nodes in Figure 1. For Gaussian responses it reduces to a Gibbs sampler. The full conditionals of the parameters w , τ_j^2 , γ_j , $j = 1, \dots, p$, and the means $\mathbf{m} = (m_1, \dots, m_l, \dots, m_q)'$ of the conditionally Gaussian variables $\xi|m_l \sim N(m_l, 1)$, $m_l = \pm 1$, are available in closed form and are included in Algorithm 1 in the appendix. They do not depend on the specific exponential family chosen for the responses.

The full conditionals for $\alpha = (\alpha_1, \dots, \alpha_p)'$ and $\xi = (\xi_1', \dots, \xi_p')'$ depend on the “conditional” design matrices $\mathbf{X}_\alpha = \mathbf{X} \text{blockdiag}(\xi_1, \dots, \xi_p)$ and $\mathbf{X}_\xi = \mathbf{X} \text{diag}(\text{blockdiag}(\mathbf{1}_{d_1}, \dots, \mathbf{1}_{d_p})\alpha)$, respectively, where $\mathbf{1}_d$ is a $d \times 1$ vector of ones and $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_p)$ is the concatenated design matrix. For Gaussian responses, the full conditionals are given by

$$\begin{aligned} \alpha|\cdot &\sim N(\mu_\alpha, \Sigma_\alpha) \text{ with} \\ \Sigma_\alpha &= \left(\frac{1}{\sigma^2} \mathbf{X}_\alpha^T \mathbf{X}_\alpha + \text{diag}(\gamma \tau^2)^{-1} \right)^{-1}, \mu_j = \frac{1}{\sigma^2} \Sigma_\alpha \mathbf{X}_\alpha^T \mathbf{y}, \text{ and} \\ \xi|\cdot &\sim N(\mu_\xi, \Sigma_\xi) \text{ with} \\ \Sigma_\xi &= \left(\frac{1}{\sigma^2} \mathbf{X}_\xi^T \mathbf{X}_\xi + \mathbf{I} \right)^{-1}; \mu_j = \Sigma_\xi \left(\frac{1}{\sigma^2} \mathbf{X}_\xi^T \mathbf{y} + \mathbf{m} \right). \end{aligned} \tag{5}$$

For non-Gaussian responses, we use an MH algorithm with a penalized IWLS proposal (P-IWLS) based on an approximation of the current posterior mode, described in detail in Brezger and Lang (2006) (Sampling scheme 1, Section 3.1.1). The MH step uses a Gaussian (i.e. second order Taylor) approximation around the approximate mode of the full conditional as its proposal distribution. To decrease computational complexity, we modify the IWLS algorithm by using the mean of the proposal distribution of the previous step instead of the posterior mode. Because of the prohibitive computational cost for large q and p (and low acceptance rates for non-Gaussian response for high-dimensional IWLS proposals), neither α nor ξ are updated all at once. Rather, both α and ξ are

split into b_α (b_{ξ}) update blocks that are updated sequentially conditional on the states of all other parameters. For efficient and numerically stable draws from the multivariate Gaussian densities in (5) or in the IWLS algorithm, we use QR decompositions of the covariance matrices. Alternatively, Cholesky decompositions as in Lang and Brezger (2004) or Brezger and Lang (2006) may be employed.

After updating the entire α - and ξ -vectors, each subvector ξ_j is rescaled so that $|\xi_j|$ has mean 1, and the associated α_j is rescaled accordingly so that $\beta_j = \alpha_j \xi_j$ is unchanged:

$$\xi_j \rightarrow \frac{d_j}{\sum_i |\xi_{ji}|} \xi_j \quad \text{and} \quad \alpha_j \rightarrow \frac{\sum_i |\xi_{ji}|}{d_j} \alpha_j.$$

This rescaling is advantageous since α_j and ξ_j are not identifiable and thus their sampling paths can wander off into extreme regions of the parameter space without affecting the fit, e.g. α_j becoming extremely large while entries in ξ_j simultaneously become extremely small. By rescaling, we retain the interpretation of α_j as a scaling factor representing the importance of the model term associated with it and avoid numerical problems that can occur for extreme parameter values.

By default, starting values $\beta^{(0)}$ are drawn randomly in three steps: First, we do 5 Fisher scoring steps with fixed, large hypervariances. Second, for each chain run in parallel, Gaussian noise is added to this vector, and third its constituting p subvectors are scaled with variance parameters $\gamma_j \tau_j^2$ ($j = 1, \dots, p$) drawn from their priors. This means some of the p model terms are set close to zero initially, and the remainder is in the vicinity of their respective ridge-penalized maximum likelihood estimates. Starting values for $\alpha^{(0)}$ and $\xi^{(0)}$ are then computed via $\alpha_j^{(0)} = d_j^{-1} \sum_i |\beta_{ji}^{(0)}|$ and $\xi_j^{(0)} = \beta_j^{(0)} / \alpha_j^{(0)}$.

Function selection, i.e. selection of coefficient blocks $j = 1, \dots, p$ can be based on the posterior inclusion probabilities $P(\gamma_j = 1 | \mathbf{y})$. Instead of estimating them simply through the proportion of MCMC samples (t) for which $\gamma_j^{(t)} = 1$, which may have high sampling variance, we improve precision through the Rao-Blackwellized estimate $\hat{P}(\gamma_j = 1 | \mathbf{y}) = T^{-1} \sum_t P(\gamma_j^{(t)} = 1 | \cdot)$. The full conditional probabilities $P(\gamma_j^{(t)} = 1 | \cdot)$ are directly available from step 14 of the MCMC Algorithm 1 (see appendix). Based on these quantities, we can

evaluate the evidence for the effect of a continuous covariate being linear or nonlinear, whether a model term has a relevant effect at all, which interaction terms are relevant, and so on.

3. EMPIRICAL EVALUATION

3.1. Simulations

To evaluate the performance of the peNMIG prior for function selection in generalized additive models, we conducted extensive simulations representing scenarios of different complexity comprising different response types, sample sizes, signal to noise ratios, correlated and uncorrelated covariates, varying degrees of concurvity, and predictors of either high or low sparsity. As competitors, we considered component-wise boosting (Hothorn et al., 2010), ACOSSO (Storlie et al., 2010), as well as SPAM (Ravikumar et al., 2009), the double shrinkage GAM proposed by Marra and Wood (2011), the HGAM approach of Meier et al. (2009) and finally Bayesian additive regression trees (Chipman et al., 2010) as a non-parametric, “black-box prediction” alternative. As benchmark, we used a conventional GAM based on the true model structure. Note that ACOSSO, SPAM and HGAM are available for Gaussian responses only. Predictive deviance and the ability to recover the correct model complexity were used as performance measures. Detailed description and graphical summaries of these simulations can be found in Section C of the appendix. Extensive additional simulation studies are described in the first author’s dissertation (Scheipl, 2011a).

The main conclusion that can be drawn from the simulations is that the proposed approach is highly competitive to previous suggestions while having the advantage of being applicable both for generalized exponential family regression (while most previous suggestions are restricted to Gaussian responses) and a much broader class of model terms (most previous suggestions are restricted to univariate smooth functions). Estimation based on the peNMIG prior typically achieves good estimation performance simultaneously with high accuracy in determining the correct model specification. It is robust

against correlated covariates and low signal-to-noise ratios. Prediction accuracy is robust against concurvity, however very strong concurvity of course leads to diminished selection accuracy. Selection of large coefficient blocks such as random effects for non-Gaussian response can be problematic: For both Poisson and binary responses, the selection accuracy in this case was very low, albeit without adverse effects on the estimation of the coefficients.

Robustness to hyperparameter settings was investigated by running the simulations for combinations of $v_0 = 0.01, 0.005, 0.00025$ and $(a_\tau, b_\tau) = (5, 25), (5, 50), (10, 30)$. Prediction accuracy was very robust against different hyperparameter configurations in all the settings we considered while variable selection and model choice were more sensitive to varying hyperparameters, because estimated posterior inclusion probabilities for small effects are sensitive to the value for v_0 . This parameter controls the threshold of relevance of the model terms: in general, very small v_0 means small effects are more likely to be included in the model, while larger v_0 yield more conservative selection properties. However, the conducted simulations and the applications presented in the following sections provide a solid foundation for the choice of appropriate values for a wide range of applied problems. We have successfully fitted all of the application examples and the binary classification data discussed in the following section with a default prior that uses $v_0 = 0.00025, (a_\tau, b_\tau) = (5, 25)$ and $(a_w, b_w) = (1, 1)$.

3.2. Binary Classification Benchmarks

We use a collection of 21 data sets for binary classification to investigate the performance of our approach on some well known benchmarks. The same collection has previously been used for benchmarking in Meyer et al. (2003) which contains some more details on the datasets that we use. Table 1 gives an overview of the datasets and their characteristics. We evaluate prediction performance based on the deviance values for a 20-fold cross validation on each dataset. Predictive deviance \bar{D} is defined as twice the average negative log likelihood $\bar{D} = -2/n_P \sum_{i=1}^{n_P} L(y_{P,i}, \hat{\eta}_{P,i})$ in the test sample where y_P and $\hat{\eta}_P$ are the out-of-sample responses and the posterior mean of the linear predictor for the test sample.

data set	n	covariates	of which factors	balance
BreastCancer	683	9	9	0.54
Cards	653	15	5	0.83
Circle	1200	2	0	0.97
Heart1	296	13	5	0.85
HouseVotes84	232	16	0	0.87
Ionosphere	351	33	0	0.56
PimaDiab	768	8	0	0.54
Sonar	208	60	0	0.87
Spirals	1200	2	0	1.00
chess	3196	36	1	0.91
credit	1000	24	9	0.43
hepatitis	80	19	0	0.19
liver	345	6	0	0.72
monks3	554	6	4	0.92
musk	476	166	0	0.77
promotergene	106	57	57	1.00
ringnorm	1200	20	0	1.00
threenorm	1200	20	0	1.00
tictactoe	958	9	9	0.53
titanic	2201	3	1	0.48
twonorm	1200	20	0	1.00

Table 1: Characteristics of UCI data sets. “Balance” is the ratio between the number of observations in the larger class and the number of observations in the smaller class, i.e. it is 1 if the data set is balanced.

The size of the test sample is n_p . As for the experiments with simulated data, we use component-wise boosting with separate base learners for the linear and smooth parts of covariate influence and compare prediction performance of the boosting models to our approach. For boosting, we determine the stopping iteration m_{stop} based on the out-of-bag risk in 10 bootstrap samples of the training sample.

The second metric we are interested in is the parsimony of the estimated models. We simply count the number of model terms (or baselearners for boosting) included in the model, i.e., we count the model terms with marginal posterior inclusion probabilities greater than 0.5. For boosting, we count a baselearner as included if it was selected at least once before iteration m_{stop} in more than half of the bootstrap samples used to determine m_{stop} .

The data are preprocessed in order to preempt numerical problems: All covariates

with less than six unique values are treated as factor variables. All numeric covariates are logarithmized if their absolute skewness is larger than two and standardized to have mean zero and unit standard deviation. Incomplete observations are removed. All numeric covariates are associated with both a linear and a smooth effect. For data sets `credit`, `Cards`, `Heart1`, `Ionosphere`, `hepatitis`, `Sonar` and `musk`, we use NMIG instead of peNMIG for terms with $d = 1$ (i.e. linear terms and binary factors) to reduce the posterior’s dimensionality. Estimates are based on samples from eight parallel chains with a burn-in of 1000 iterations, followed by a sampling phase of 5000 iterations of which we save every fifth. We report results for $(a_\tau, b_\tau) = (5, 25)$, $w \sim \text{Beta}(1, 1)$ and $v_0 \in \{0.005, 0.00025\}$. Results with $(a_\tau, b_\tau) = (5, 50)$ and $w \sim \text{Beta}(20, 40)$ were qualitatively similar and are omitted for clarity of presentation. An unabridged description is in Scheipl (2011a, Ch. 4.2).

Figure 3 shows the prediction accuracy for the 21 datasets. Most outliers with large deviances are due to the sampler getting stuck for some of the parallel chains in specific folds for some of the data sets. By rerunning the analysis with different starting values or random seeds or manual postprocessing of the posterior samples, these presumably could have been avoided. We include them unchanged to provide a more realistic picture of the reliability of our approach. Practicioners should always check acceptance rates and convergence diagnostics when using MCMC-based methods. Note that the predictive performance of our approach is usually more variable than that achieved by `mboost` but has lower median predictive deviances in all of the datasets for all prior specifications (Results for $(a_\tau, b_\tau) = (5, 50)$ and $w \sim \text{Beta}(20, 40)$ not shown). Predictive performance is very robust against different hyperparameter settings, even for large p/n ratio where the influence of the hyperparameters on the posterior is relatively stronger. To investigate the parsimony of the fitted models, i.e. whether equivalent or better prediction can be achieved by simpler models, we plot the differences in predictive deviances versus the difference in the proportion of potential model terms included in the models in Figure 4 (Results for `Spirals`, `tictactoe` and `titanic` not shown because there were no differences in sparsity). Positive values on the vertical axis indicate smaller deviance for

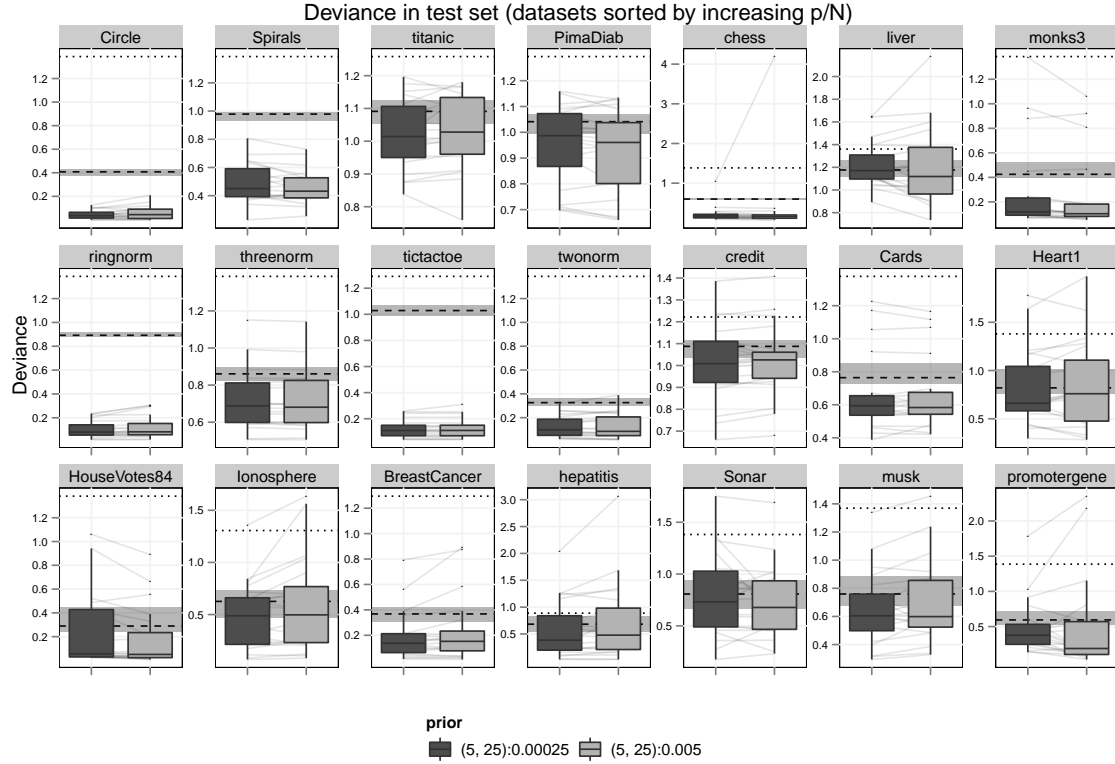


Figure 3: UCI data I: Predictive Deviances for 20-fold crossvalidation. Box-plots show results for the different prior settings, the horizontal ribbon shows results for mboost: shaded region gives IQR, dashed line represents median. Dark grey lines connect results for the same fold. Dotted line gives predictive deviance of the null model on the full data set.

our approach, and positive values on the horizontal axis indicate a sparser fit for our approach. Figure 4 shows that our approach achieves its relatively more precise predictions with smaller models on the large majority of the benchmark data sets. The only exceptions are datasets chess, where the increased precision is achieved at the cost of less sparse models, and, to a much lesser extent, twonorm and threenorm. Neither absolute performance nor performance relative to boosting seems to be tied to any of the easily observable characteristics of the data sets (i.e. p , n , p/n , or balancedness). No clear picture emerges for the differences between the prior specifications: As expected, a smaller value for v_0 tends to yield larger models, cf. datasets chess, twonorm, Ionosphere, musk, but there are counterexamples as well, e.g. threenorm. Both predictive deviance and sparsity results are more sensitive towards v_0 than towards (a_τ, b_τ) (Results for $(a_\tau, b_\tau) = (5, 50)$)

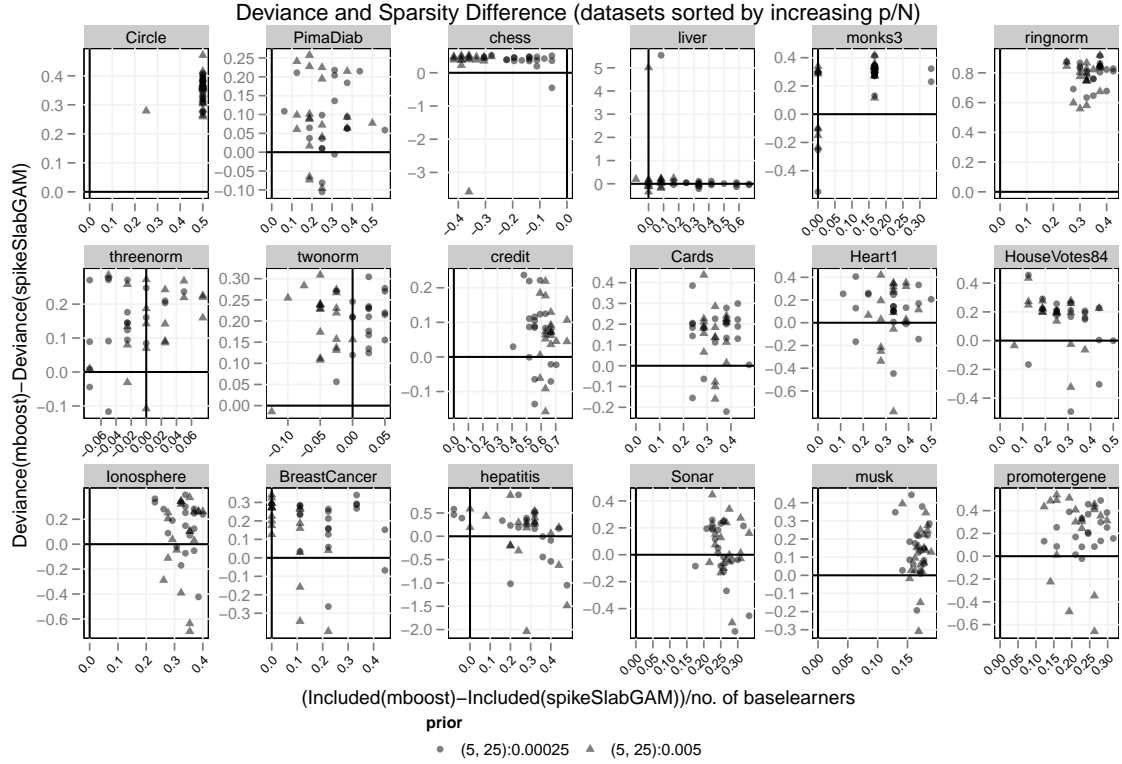


Figure 4: UCI data I: Difference in proportion of included model terms versus differences in predictive deviances. Positive values denote smaller deviances / models for our approach compared to mboost. Results for Spirals, titanic and tictactoe not shown because there were no differences in sparsity.

not shown). Using an informative prior $w \sim \text{Beta}(20, 40)$ to enforce model sparsity has no appreciable effect and does not influence prediction quality (Results not shown).

More generally, the performance of our approach shows that it is very competitive to componentwise boosting and that neither relative nor absolute performance deteriorate in very high-dimensional problems (cf. results for musk with $n = 476$ and 332 potential model terms, of which 166 are smooth terms.).

3.3. Case Study: Survival of Surgical Patients with Severe Sepsis

Data: We use data on the survival of 462 patients with severe sepsis that was collected in the intensive care unit of the Department of Surgery at Munich's Großhadern hospital

between March 1, 1993, and February 28, 2005. Hofner et al. (2011) have previously analysed this data set. The follow-up period was 90 days after the beginning of intensive care, with one drop-out after 66 days and 179 patients surviving the observation period.

Models: We use a piecewise exponential model (PEM) (Fahrmeir and Tutz, 2001, Ch. 9) to model the hazard rate $\lambda(t, \mathbf{x})$ of the underlying disease process, i.e. for fixed time intervals defined by cutpoints $\kappa = (\kappa_0 = 0, \kappa_1, \dots, \kappa_L = t_{\max})$, where t_{\max} is the maximal follow-up time, the hazard rate for subject i at time t , $\kappa_{j-1} < t \leq \kappa_j$ in the j^{th} interval is given by

$$\lambda(t, \mathbf{x}_i) = \exp \left(g_0(j) + \sum_{l=1}^L g_l(j) v_{il}(j) + \sum_{m=1}^M f_m(u_{im}(j)) + \mathbf{z}_i(j)' \boldsymbol{\gamma} \right)$$

where $g_0(j)$ represents the baseline hazard rate in interval j , $g_l(j)$, $l = 1, \dots, L$, are time-varying effects of covariates $v_{il}(j)$, $f_m(u_{im}(j))$, $j = 1, \dots, J$, are nonlinear effects of continuous covariates $u_{im}(j)$ and $\mathbf{z}_i(j)' \boldsymbol{\gamma}$ contains linear, parametric effects. All time-dependent quantities are assumed to be piecewise constant on the intervals such that e.g. $g_0(t) = g_0(j)$ for all $t \in (\kappa_{j-1}, \kappa_j]$. The interval borders $\kappa = (0, 5, 15, 25, \dots, 85, 90)$ were chosen based on the shape of a nonparametric estimate of the marginal hazard rate. The likelihood for the PEM is equivalent to that of a Poisson model with (1) one observation for each interval for each subject, yielding 2826 pseudo-observations in total, (2) offsets $o_{ij} = \max(0, \min(\kappa_j - \kappa_{j-1}, t_i - \kappa_{j-1}))$, where t_i is the observed time under risk for subject i and (3) responses y_{ij} equal to the event indicators δ_{ij} , with $\delta_{ij} = 0$ if subject i survived interval j and $\delta_{ij} = 1$ if not.

Our aim is twofold: We want to (1) estimate a model that allows assessment of the influence of each available covariate on the prognosis of patients, accounting for possibly time-varying and/or nonlinear effects and (2) use this setting to evaluate the stability of the selection and estimation of increasingly complex models on real data. Important covariates included in the analysis are for example the age of the patient, the haemoglobin concentration, the presence of a fungal infection, different types of operations and the Apache II score (a measure for the severity of disease), see Hofner et al. (2011) for a

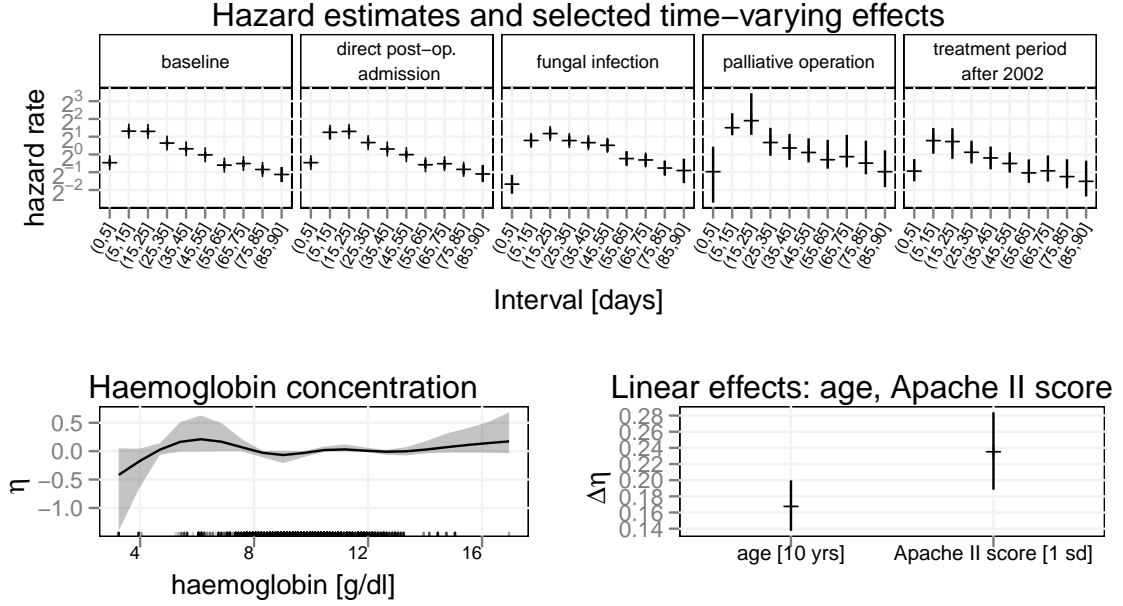


Figure 5: Posterior means of effects with (pointwise) 80% credible intervals. Top: Baseline hazard rate and baseline hazard rate plus the time-varying and time-constant effects for direct postoperative admission, presence of a fungal infection, palliative operation and beginning of treatment after 2002. Bottom: smooth effect of haemoglobin concentration and linear effects of age (10 year increase) and Apache II score, a measure for disease severity (increase of score by 1 standard deviation).

complete description.

Full Data Results We perform term selection for a maximal model which includes the (linear and non-linear) effects of all 20 covariates as well as their time-varying effects, i.e. 48 potential model terms with 262 coefficients in total. Hyperparameters were set to the default values determined in the simulation studies, i.e. $a_\tau = 5, b_\tau = 25, v_0 = 0.00025$ and $a_w = b_w = 1$. Estimates are based on 8 parallel chains running for 20000 iterations each after a burn-in of 500 iterations, with every 10^{th} iteration saved. We use a first order random walk prior for the log-baseline and the time-varying effects to regularize their roughness, i.e., we use an intrinsic GMRF prior (on the line) for the piecewise constant time-varying quantities (denoted as $MRF(Interval)$ in the following).

The estimated marginal inclusion probabilities indicate a fairly sparse model, with posterior marginal inclusion probabilities greater than 0.10 for only 10 terms, as shown in

Term	$P(\gamma = 1)$
MRF(Interval)	1.00
palliative operation	0.19
treatment period	0.71
Age, linear	0.99
Apache II score, linear	1.00
Haemoglobin concentration, smooth	0.38
MRF(Interval):direct postoperative admission	0.28
MRF(Interval):fungal infection	1.00
MRF(Interval):palliative operation	0.38
MRF(Interval):treatment period	0.13

Table 2: Posterior means of marginal inclusion probabilities $P(\gamma = 1)$ (only given for terms with $P(\gamma = 1) > .1$).

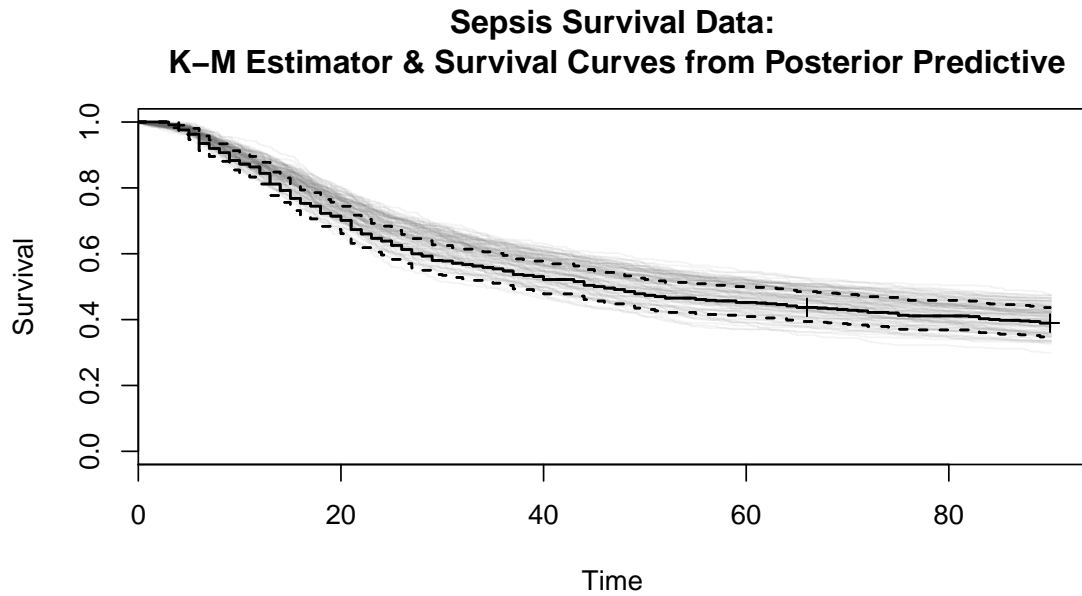


Figure 6: Kaplan-Meier estimate of the survival curve for observed data in black. Grey overlays are survival curves for 100 replicates of survival time vectors generated from the posterior predictive distribution.

Table 2. The estimated effects for this subset of terms are visualized in Figure 5. To verify the suitability of the model, we perform a posterior predictive check and generate 100 replicates of survival times from the posterior predictive distribution. Figure 6 indicates that the fit is satisfactory, although there seems to be a tendency to overestimate survival rates until about day 70.

Predictive Performance Comparison We subsample the data 20 times to construct independent training data sets with 415 patients each and test data sets with the remaining 47 patients to evaluate the precision of the resulting predictions and compare predictive performance to that of equivalent component-wise boosting models fitted with `mboost`. Results for our approach are based on 8 parallel chains, each running for 5000 iterations after 500 iterations of burn-in, with every fifth iteration saved. Component-wise boosting results are based on a stopping parameter determined by a 25-fold bootstrap of the training data, with a maximal iteration number of 1500.

The previous analysis by Hofner et al. (2011) has used expert knowledge to define a set of six covariates forced into the model (indicators for presence of malignant primary disease, palliative operation and beginning of treatment after 2002, as well as sex, age and Apache II score). We compare results for four model specifications of increasing complexity that suggest themselves: a model with only the main effects of the pre-selected covariate set, a model with main effects and time-varying effects for the pre-selected covariate set, a model with main effects for all 20 covariates and the model with main effects and time-varying effects for all 20 covariates which was applied to the whole data set (see above). As in the previous section, main effects for numerical covariates such as age were split into linear and non-linear parts. Figure 7 shows the predictive deviances achieved by the different model specifications. Predictive deviance is defined as $-2 \sum_i^{N_t} \sum_j^{J(i)} \delta_{ij} (\log(\hat{\lambda}_j) + \hat{\eta}_{ij}) - o_{ij} \hat{\lambda}_j \exp(\hat{\eta}_{ij})$, where $i = 1, \dots, N_t$ indicates the subjects in the test set and $j = 1, \dots, J(i)$ indicates the intervals in which individual i was under risk, $\hat{\lambda}_j$ and $\hat{\eta}_{ij}$ are the respective posterior predictive means. For this data set, models with higher maximal complexity seem to offer no relevant improvement in terms of prediction accuracy compared to the simplest model based only on the pre-selected covariate set without time-varying effects. Most of the models yield essentially equivalent predictions. However, it is reassuring to see that the predictive performance of our approach is not degraded at all by the specification of vastly over-complex models in a setting where the underlying structure seems to be fairly simple. In contrast, prediction accuracy for component-wise boosting decreases markedly for the models including time-varying

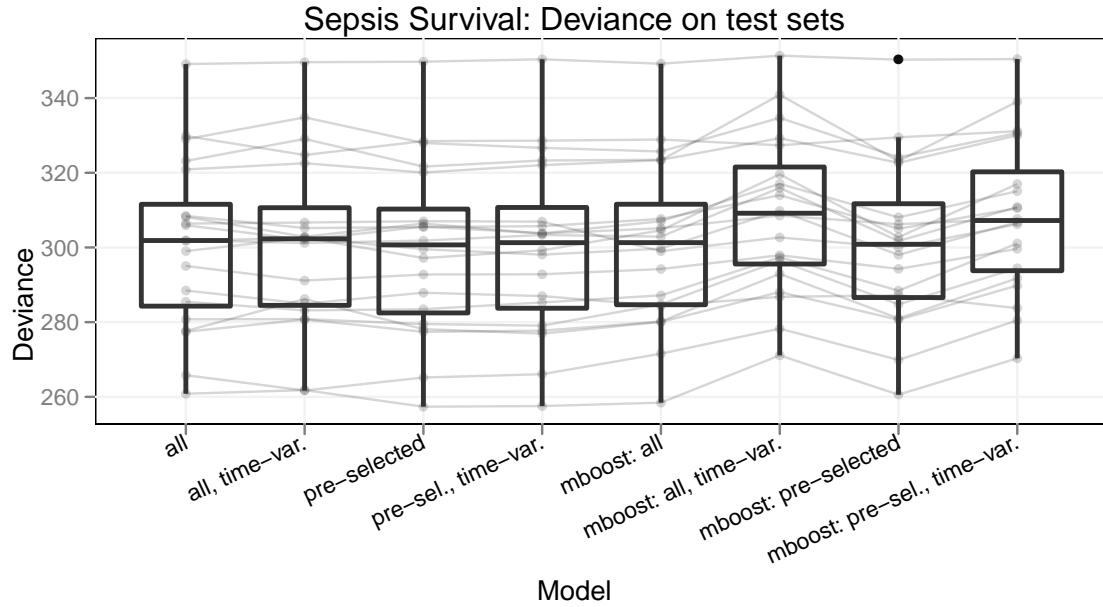


Figure 7: Predictive deviances for 20 subsampling test sets for the sepsis survival data (lower is better). Grey lines connect results from identical folds.

effects in this setting.

The stability of the marginal term inclusion probabilities across subsamples is fairly good, indicating that the term selection is robust to small changes in the data. All model specifications identified essentially the same subset of important effects from the set of pre-selected covariates (i.e., indicators for palliative operation and beginning of treatment after 2002 and linear effects of age and Apache II score), and also the same time-varying effects (i.e., time varying effects for palliative operation and beginning of treatment after 2002). Figure 8 shows the posterior means of inclusion probabilities $P(\gamma = 1)$ across 20 subsampled training data sets for each of the 4 model specifications.

Additional case studies for geoaddivitive regression of net rent levels in Munich and an additive mixed model for binary responses from a large study on hymenoptera venom allergy can be found in Section D of the appendix.

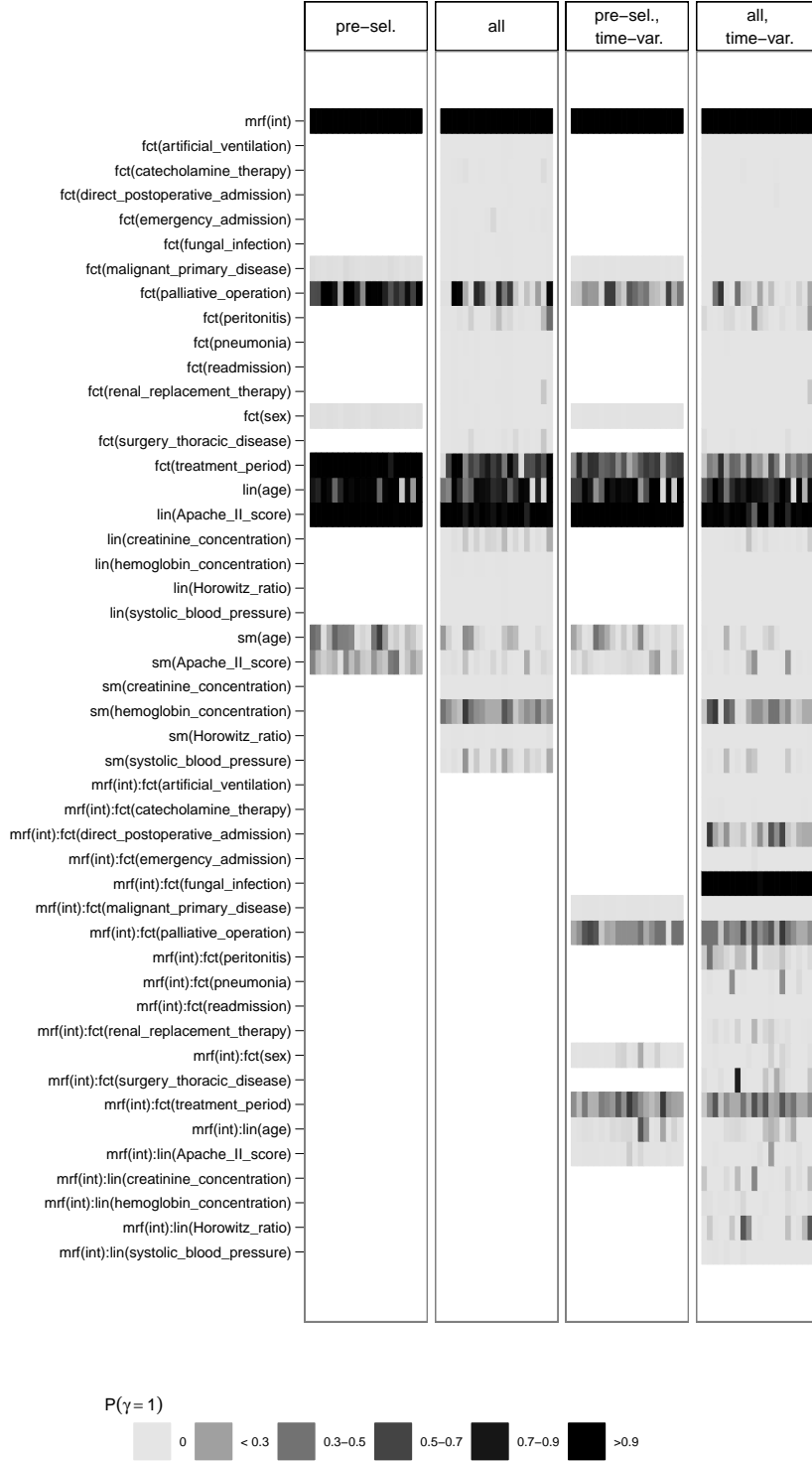


Figure 8: Posterior means of inclusion probabilities $P(\gamma = 1)$ across 20 sub-sampled training data sets for the 4 model specifications.

4. CONCLUSIONS

In this paper, we have proposed a general Bayesian framework for conducting function selection in exponential family structured additive regression models. Inspired by stochastic search variable selection approaches and the general idea of spike-and-slab priors, we introduced a non-identifiable multiplicative parameter expansion where selection or deletion of coefficient batches (such as parameters representing a spline basis or random intercepts) is associated with a scalar scaling factor only. This reparameterization alleviates the notorious mixing problems that would appear in a naive implementation of our prior structure.

The main advantages of the proposed peNMIG prior structure are (1) its general applicability for various types of responses (in particular non-Gaussian responses), (2) the availability of a supplementing R package that makes all methods immediately accessible and reproducible, and (3) the good performance demonstrated in simulations and applications with fairly low sensitivity with respect to hyperparameter settings and substantiated in theoretical investigations of the shrinkage properties of peNMIG. The class of models that can be fitted in this framework can be extended fairly easily by considering other latent Gaussian or latent exponential family models that can be implemented via data augmentation.

A different perspective on stochastic search variable selection approaches is to consider them as a possibility for implementing Bayesian model averaging. Since so far most applications and implementations of Bayesian model averaging are restricted to linear or generalized linear models, our approach offers a necessary extension of Bayesian model averaging implementations to a much broader model class. As shown in Section 3.1, it offers improved prediction accuracy and allows for a principled inclusion of uncertainty about term selection and model structure into inferential statements.

COMPUTATIONAL DETAILS

The approach described and evaluated in this paper is implemented in the R-package `spikeSlabGAM` (Scheipl, 2011b).

ACKNOWLEDGEMENTS

We are indebted to Franziska Ru  ff for letting us use the insect allergy data set as an application example and to Wolfgang Hartl for letting us use the sepsis survival data. Financial support from the German Science Foundation (grant FA 128/5-1) is gratefully acknowledged. We thank two referees for their constructive comments which helped to substantially improve the paper.

A. MCMC algorithm

B. Problems of the Conventional NMIG Prior when Selecting Coefficient Blocks

Previous approaches for Bayesian variable selection have primarily concentrated on selection of single coefficients (George and McCulloch, 1993; Ishwaran and Rao, 2005) or used very low dimensional bases for the representation of smooth effects. E.g. Cottet et al. (2008) use a pseudo-spline representation of their cubic smoothing spline bases with only 3 to 4 basis functions. In the following, we argue that conventional blockwise Gibbs sampling is ill suited for updating the state of the Markov chain when sampling from the posterior of an NMIG model even for moderately large coefficient blocks. We show that mixing for γ_j will be very slow for blocks of coefficients β_j with $d_j \gg 1$. We suppress the index j in the following.

The following analysis shows that, even if the blockwise sampler is initially in an ideal

Algorithm 1 MCMC sampler for peNMIG

- 1 Initialize $\tau^{2(0)}, \gamma^{(0)}, \sigma^{2(0)}, w^{(0)}$ and $\beta^{(0)}$ ($\beta^{(0)}$ via IWLS for non-Gaussian response)
 - 2 Compute $\alpha^{(0)}, \xi^{(0)}, X_\alpha^{(0)}$
 - 3 **for** iterations $t = 1, \dots, T$ **do**
 - 4 **for** blocks $b = 1, \dots, b_\alpha$ **do**
 - 5 update $\alpha_b^{(t)}$ from its fcd (Gaussian case, see (5))/ via P-IWLS
 - 6 set $X_\xi^{(t)} = X \text{diag}(\text{blockdiag}(\mathbf{1}_{d_1}, \dots, \mathbf{1}_{d_p})\alpha^{(t)})$
 - 7 update $m^{(t)}$ from their fcd: $P(m_l^{(t)} = 1|\cdot) = \frac{1}{1+\exp(-2\xi_l^{(t)})}$, $l = 1, \dots, q$
 - 8 **for** blocks $b = 1, \dots, b_\xi$ **do**
 - 9 update $\xi_b^{(t)}$ from its fcd (Gaussian case, see (5))/ via P-IWLS
 - 10 **for** model terms $j = 1, \dots, p$ **do**
 - 11 rescale $\xi_j^{(t)}$ and $\alpha_j^{(t)}$
 - 12 set $X_\alpha^{(t)} = X \text{blockdiag}(\xi_1^{(t)}, \dots, \xi_p^{(t)})$
 - 13 update $\tau_1^{2(t)}, \dots, \tau_p^{2(t)}$ from their fcd: $\tau_j^{2(t)}|\cdot \sim \Gamma^{-1}\left(a_\tau + 1/2, b_\tau + \frac{\alpha_j^{2(t)}}{2\gamma_j^{(t)}}\right)$
 - 14 update $\gamma_1^{(t)}, \dots, \gamma_p^{(t)}$ from their fcd: $\frac{P(\gamma_j^{(t)}=1|\cdot)}{P(\gamma_j^{(t)}=v_0|\cdot)} = v_0^{1/2} \exp\left(\frac{(1-v_0)}{2v_0} \frac{\alpha_j^{2(t)}}{\tau_j^{2(t)}}\right)$
 - 15 update $w^{(t)}$ from its fcd:
 $w^{(t)}|\cdot \sim \text{Beta}\left(a_w + \sum_j^p \delta_1(\gamma_j^{(t)}), b_w + \sum_j^p \delta_{v_0}(\gamma_j^{(t)})\right)$
 - 16 **if** y is Gaussian **then**
 - 17 update $\sigma^{2(t)}$ from its fcd: $\sigma^{2(t)}|\cdot \sim \Gamma^{-1}\left(a_{\sigma^2} + n/2, b_{\sigma^2} + \frac{\sum_i^n (y_i - \eta_i^{(t)})^2}{2}\right)$
-

state for switching between the spike and the slab parts of the prior, i.e. a parameter constellation so that the full conditional probability $P(\gamma = 1|\cdot) = .5$, such a switch is very unlikely in subsequent iterations for coefficient vectors with more than a few entries given the NMIG prior hierarchy.

Assume that the sampler starts out in iteration (0) with a parameter configuration of $a_t, b_t, v_0, w, \tau_{(0)}^2$ and $\beta_{(0)}$ so that $P(\gamma_{(0)} = 1|\cdot) = .5$. We set $w = .5$. The parameters for which $P(\gamma = 1|\cdot) = .5$ satisfy the following relations:

$$\frac{P(\gamma = 1|\cdot)}{P(\gamma = v_0|\cdot)} = v_0^{d/2} \exp\left(\frac{(1-v_0)}{2v_0} \frac{\sum^d \beta^2}{\tau^2}\right) = 1,$$

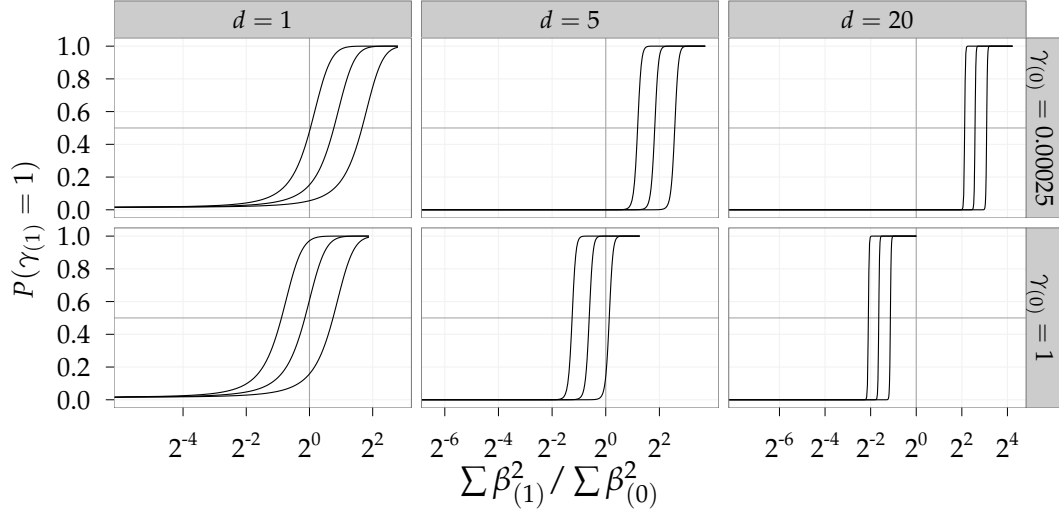


Figure 9: $P(\gamma)$ as a function of the relative change in $\sum^d \beta^2$ for varying $d, \gamma_{(0)}$: Inclusion probability in iteration (1) as a function of the ratio between the sum of squared coefficients in iteration (1) and (0). Lines in each panel correspond to $\tau_{(1)}^2$ equal to the median of its full conditional and the .1- and .9-quantiles. Upper row is for $\gamma_{(0)} = 1$, lower row for $\gamma_{(0)} = v_0$. Columns correspond to $d = 1, 5, 20$. Fat gray grid lines denote inclusion probability = .5 and ratio of coefficient sum of squares = 1

so that $P(\gamma = 1|\cdot) > .5$ if

$$\frac{\sum^d \beta^2}{d\tau^2} > -\frac{v_0}{1-v_0} \log(v_0).$$

Assuming a given value $\tau_{(0)}^2$, set

$$\sum^d \beta_{(0)}^2 = -\frac{dv_0}{1-v_0} \log(v_0) \tau_{(0)}^2.$$

Now $\gamma_{(0)}$ takes on both values v_0 and 1 with equal probability, conditional on all other parameters.

In the following iteration, $\tau_{(1)}^2$ is drawn from its full conditional $\Gamma^{-1}(a_t + d/2, b_t + \frac{\sum^d \beta_{(0)}^2}{2\gamma_{(0)}})$. Figure 9 shows $P(\gamma_{(1)} = 1|\tau_{(1)}^2, \sum^d \beta_{(1)}^2)$ as a function of $\sum^d \beta_{(1)}^2 / \sum^d \beta_{(0)}^2$ for various values of d . The 3 lines in each panel correspond to $P(\gamma_{(1)} = 1|\tau_{(1)}^2, \sum^d \beta_{(1)}^2)$ for values of $\tau_{(1)}^2$ equal to the median of its full conditional as well as the .1- and .9-quantiles. The

lower row in the Figure plots the function for $\gamma_{(0)} = 1$, the upper row for $\gamma_{(0)} = v_0$.

So, if we start in this “equilibrium state” we begin iteration (0) with $v_0, w, \tau_{(0)}^2$, and $\beta_{(0)}$ so that $P(\gamma_{(0)} = 1|\cdot) = 0.5$. We then determine $P(\gamma_{(1)} \neq \gamma_{(0)}|\tau_{(1)}^2, \sum^d \beta_{(1)}^2)$ as a function of $\sum^d \beta_{(1)}^2 / \sum^d \beta_{(0)}^2$ for

- various values of $\dim(\beta_j) = d$,
- $\gamma_{(0)} = 1$ and $\gamma_{(0)} = v_0$,
- $\tau_{(1)}^2$ at the .1, .5, .9-quantiles of its conditional distribution given $\beta_{(0)}, \gamma_{(0)}$.

The leftmost column in Figure 9 shows that moving between $\gamma = 1$ and $\gamma = v_0$ is easy for $d = 1$: For a large range of realistic values for $\sum^d \beta_{(1)}^2 / \sum^d \beta_{(0)}^2$, moving back to $\gamma_{(1)} = v_0$ from $\gamma_{(0)} = 1$ (lower panel) has reasonably large probability, just as moving from $\gamma_{(0)} = v_0$ to $\gamma_{(1)} = 1$ (lower panel) is fairly likely for realistic values of $\sum^d \beta_{(1)}^2 / \sum^d \beta_{(0)}^2$. For $d = 5$, however, $P(\gamma_{(1)} = 1|\cdot)$ already resembles a step function. For $d = 20$, if $\sum^d \beta_{(1)}^2 / \sum^d \beta_{(0)}^2$ is not smaller than 0.48, the probability of moving from $\gamma_{(0)} = 1$ to $\gamma_{(1)} = v_0$ (lower panel) is practically zero for 90% of the values drawn from $p(\tau_{(1)}^2|\cdot)$. However, draws of β that reduce $\sum^d \beta^2$ by more than a factor of 0.48 while $\gamma = 1$ are unlikely to occur in real data. It is also extremely unlikely to move back to $\gamma_{(1)} = 1$ when $\gamma_{(0)} = v_0$, unless $\sum^d \beta_{(1)}^2 / \sum^d \beta_{(0)}^2$ is larger than 2.9. Since the full conditional for β is very concentrated if $\gamma = v_0$, such moves are highly improbable and correspondingly the sampler is unlikely to move away from $\gamma = v_0$. Numerical values for the graphs in Figure 9 were computed for $a_\tau = 5$, $b_\tau = 50$, $v_0 = 0.00025$ but similar problems arise for all suitable hyperparameter configurations.

In summary, mixing of the indicator variables γ will be very slow for long subvectors. In experiments, we observed posterior means of $P(\gamma = 1)$ to be either ≈ 0 or ≈ 1 across a wide variety of settings, even for very long chains, largely depending on the starting values of the chains. A multiplicative parameter expansion offers a possible remedy, with the added benefit of inducing very desirable shrinkage properties for the resulting estimates as shown in the article.

C. Simulation Results

In the following Sections C.1 and C.2, we compare the performance of peNMIG in (generalized) additive models (GAMs) as implemented in package `spikeSlabGAM` (Scheipl, 2011b) to that of component-wise boosting (Hothorn et al., 2010) in terms of predictive MSE and complexity recovery. As a reference, we also fit a conventional GAM (as implemented in `mgcv` (Wood, 2008)) based on the “true” formula (i.e. a model without any of the “noise” terms), which we subsequently call the “oracle”-model. For Gaussian responses only, we also compare our results to those from ACOSSO (Storlie et al., 2010). ACOSSO is not able to fit non-Gaussian responses. Section C.3 investigates the effects of concurvity on effect estimates and term selection for our method and those of some recently proposed competitors.

We supply separate base learners for the linear and smooth parts of covariate influence for the component-wise boosting in order to compare complexity recovery between boosting and our approach. We use 10-fold cross validation on the training data to determine the optimal stopping iteration for `mboost` and count a baselearner as included in the model if it is selected in at least half of the cross-validation runs up to the stopping iteration. BIC is used to determine the tuning parameter for ACOSSO. We did not compare our approach to Reich et al. (2009), which is implemented for Gaussian responses, since the available R implementation is impractically slow – computation times are usually 15 - 30 times those of our peNMIG implementation.

For both Gaussian responses (Section C.1) and Poisson responses (Section C.2), the data generating process has the following structure:

- We define 4 functions
 - $f_1(x) = x$,
 - $f_2(x) = x + \frac{(2x-2)^2}{5.5}$,
 - $f_3(x) = -x + \pi \sin(\pi x)$,
 - $f_4(x) = 0.5x + 15\phi(2(x - .2)) - \phi(x + 0.4)$, where $\phi()$ is the standard normal density function,

which enter into the linear predictor. Note that all of them have (at least) a linear component.

- We define 2 scenarios:
 - a “low sparsity” scenario: Generate 16 covariates, 12 of which have non-zero influence. The true linear predictor is $\eta = f_1(x_1) + f_2(x_2) + f_3(x_3) + f_4(x_4) + 1.5(f_1(x_5) + f_2(x_6) + f_3(x_7) + f_4(x_8)) + 2(f_1(x_9) + f_2(x_{10}) + f_3(x_{11}) + f_4(x_{12}))$.
 - a “high sparsity” scenario: Generate 20 covariates, only 4 of which have non-zero influence and $\eta = f_1(x_1) + f_2(x_2) + f_3(x_3) + f_4(x_4)$.
- The covariates are either
 - $\stackrel{\text{i.i.d.}}{\sim} U[-2, 2]$ or
 - from an AR(1) process with correlation $\rho = 0.7$.

- We simulate 50 replications for each combination of the various settings.

We compare 9 different prior specifications arising from the combination of

- $(a_\tau, b_\tau) = (5, 25), (10, 30), (5, 50)$
- $v_0 = 0.00025, 0.005, 0.01$

Predictive MSE is evaluated on test data sets with 5000 observations. Complexity recovery, i.e. how well the different approaches select covariates with true influence on the response and remove covariates without true influence on the response is measured in terms of accuracy, defined as the number of correctly classified model terms (true positives and true negatives) divided by the total number of terms in the model. For example, the full model in the “low sparsity” scenario has 32 potential terms under selection (linear terms and basis expansions/smooth terms for each of the 16 covariates), only 21 of which are truly non-zero (the linear terms for the first 12 covariates plus the 9 basis expansions of the covariates not associated with the linear function $f_1()$). Accuracy in this scenario would then be determined as the sum of the correctly included model terms plus the correctly excluded model terms, divided by 32.

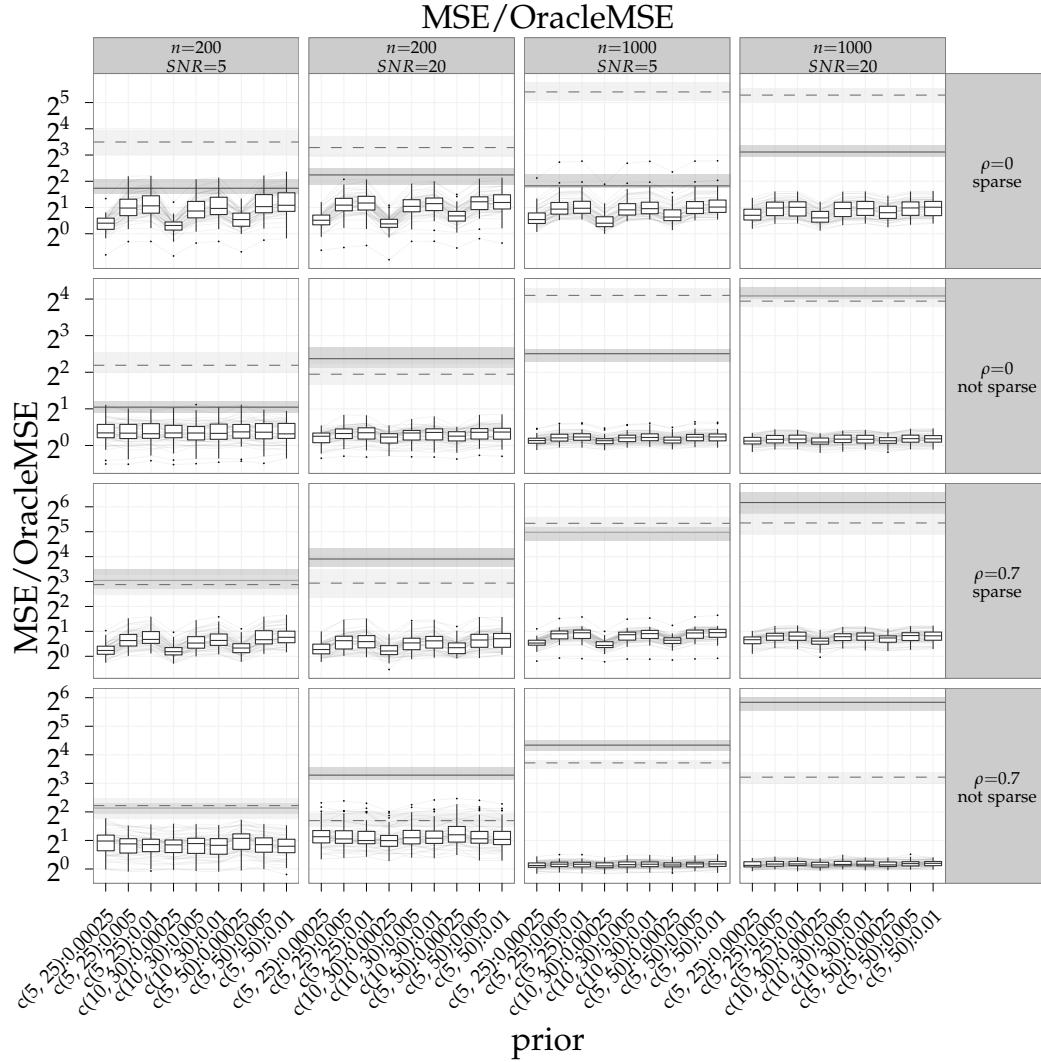


Figure 10: Prediction MSE divided by oracle MSE for Gaussian response. Boxplots show results for the different prior settings, horizontal ribbons show results for mboost (solid) and ACOSSE (dashed), respectively: Shaded region gives IQR, line represents median. Dark grey lines connect results for the same replication. Columns from left to right: 200 obs. with SNR=5, 20; 1000 obs. with SNR=5, 20. Rows from top to bottom: uncorrelated obs. with sparse and unsparse predictor, correlated obs. with sparse and unsparse predictor. Vertical axis is on binary log scale.

C.1. Gaussian response

In addition to the basic structure of the data generating process described at the beginning of this section, the data generating process for the Gaussian responses has the following

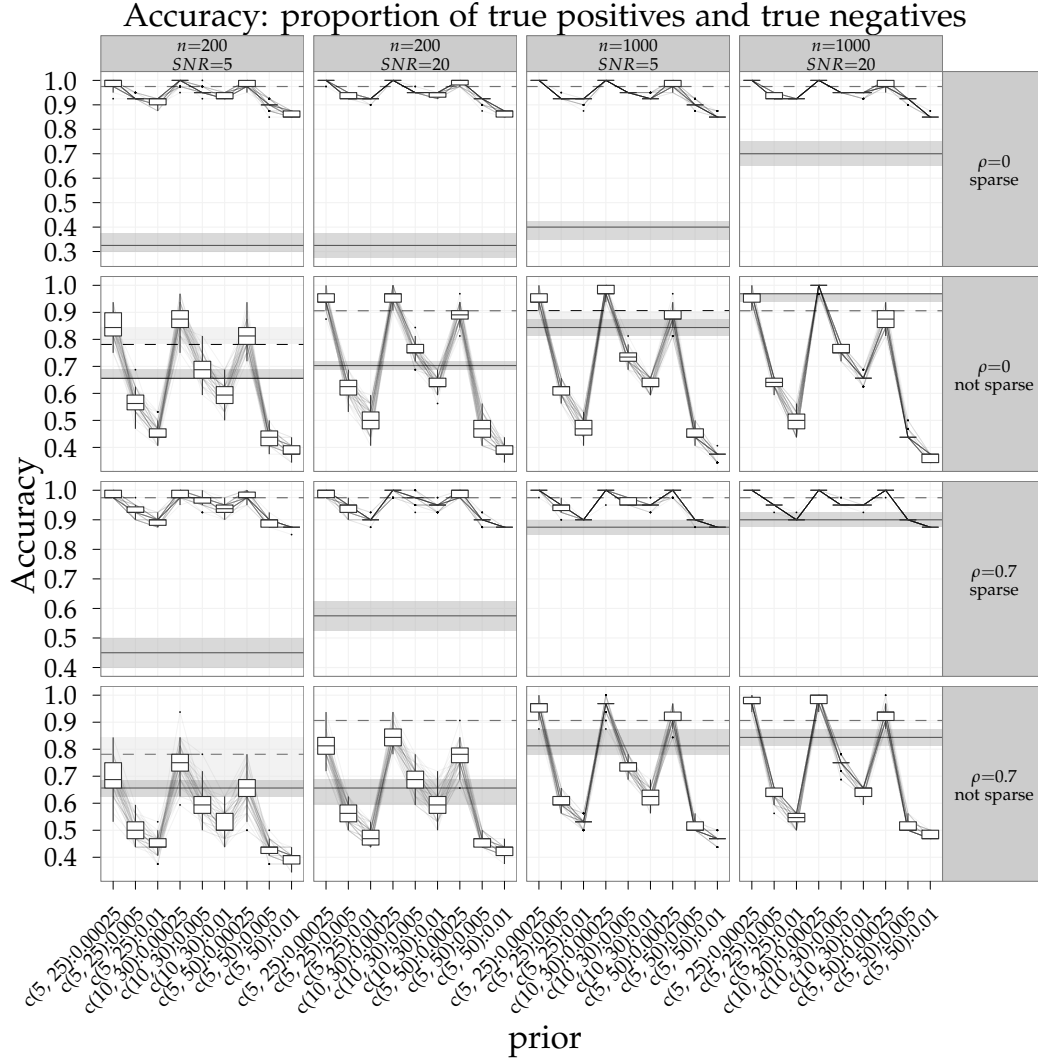


Figure 11: Complexity recovery for Gaussian response: proportion of correctly included and excluded model terms. Boxplots show results for the different prior settings, horizontal ribbons show results for mboost (solid) and ACOSSO (dashed), respectively: Shaded region gives IQR, line represents median. Dark grey lines connect results for the same replication. Columns from left to right: 200 obs. with SNR=5, 20; 1000 obs. with SNR=5, 20. Rows from top to bottom: uncorrelated obs. with sparse and unsparse predictor, correlated obs. with sparse and unsparse predictor.

properties:

- signal-to-noise-ratio SNR = 5, 20
- number of observations: $n = 200, 1000$

Figure 10 shows the mean squared prediction error divided by the one achieved by the “oracle”-model, a conventional GAM without any of the noise variables. Predictive performance is very robust against the different prior settings especially for the settings with low sparsity. Different prior settings also behave similarly within replications, as shown by the mostly parallel grey lines. Predictions are more precise than those of both boosting and ACOSSO, and this improvement in performance relative to the “true” model is especially marked for $n = 1000$ (two rightmost columns). With the exception of the first scenario, the median relative prediction MSE is < 2 everywhere, while both boosting and ACOSSO have a median relative prediction MSE above 4 in most scenarios that goes up to above 32 and 64 for ACOSSO and boosting, respectively, in the “large sample, correlated covariates” cases. In the “large sample, low sparsity” scenarios (two leftmost columns in rows two and four), the performance of our approach comes very close that of the oracle model – the relative prediction MSEs are close to one.

Figure 11 shows the proportion of correctly included and excluded terms (linear terms and basis expansions) in the estimated model. Except for $v_0 = 0.00025$, accuracy is consistently lower than for ACOSSO. However, a direct comparison with ACOSSO is not entirely appropriate because ACOSSO does not differentiate between smooth and linear terms, while mboost and our approach do. Therefore ACOSSO solves a less difficult problem. Estimated inclusion probabilities are very sensitive to v_0 and comparatively robust against (a_τ, b_τ) . Across all settings, $v_0 = 0.00025$ delivers the most precise complexity recovery, with sensitivities consistently above 0.7. The accuracy of peNMIG is better than mboost for the sparse settings (first and third rows) because the specificity of our approach is $> .97$ across settings, regardless of the prior (!), while mboost mostly achieves only very low specificity, but fairly high sensitivity.

C.2. Poisson response

In addition to the basic structure of the data generating process described at the beginning of this section, the data generating process for the Poisson responses has the following properties:

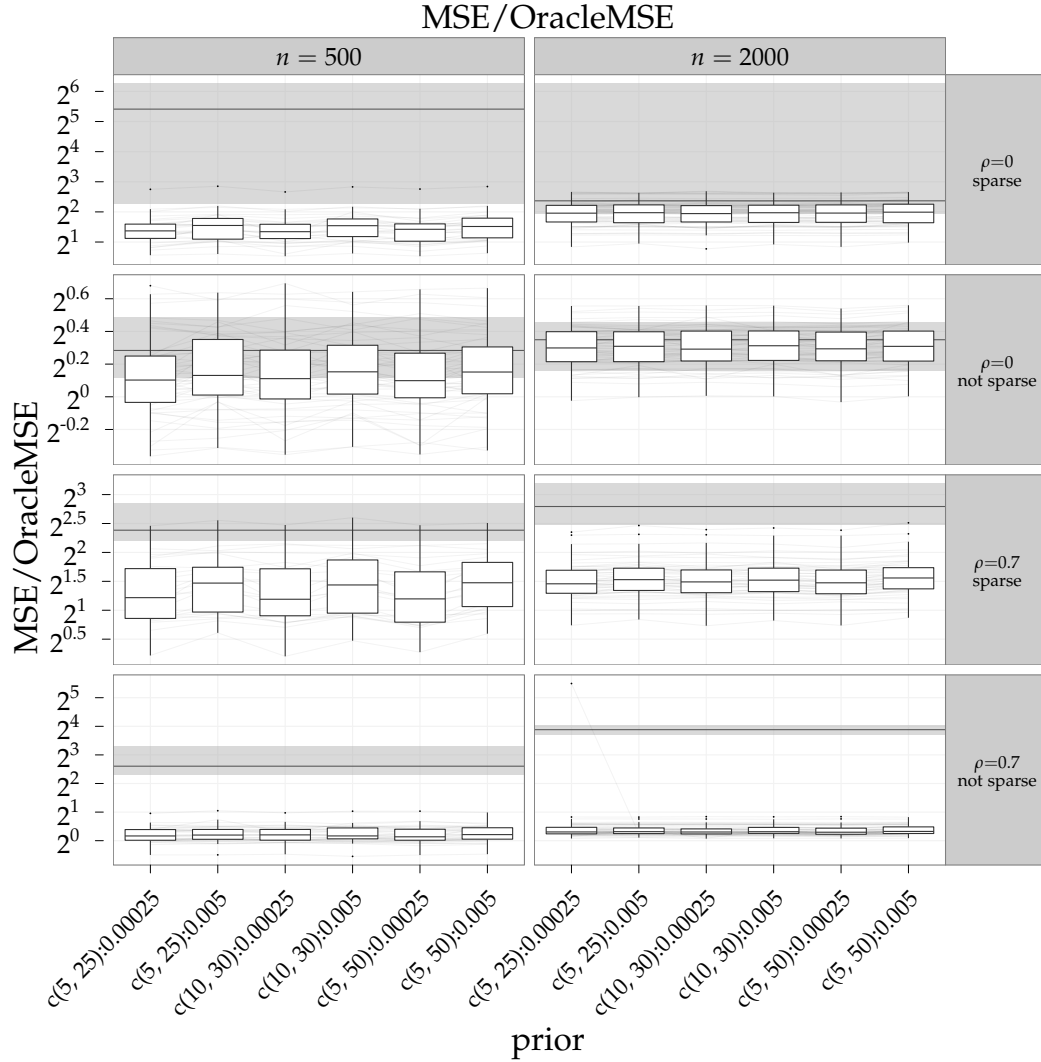


Figure 12: Prediction MSE divided by oracle MSE (on the scale of the linear predictor). Boxplots show results for the different prior settings. Horizontal ribbons show results for mboost: shaded region gives IQR, line represents median. Dark grey lines connect results for the same replication. Columns from left to right: 500 obs., 2000 obs. Rows from top to bottom: uncorrelated obs. with sparse and unsparse predictor, correlated obs. with sparse and unsparse predictor. Vertical axis is on binary log scale.

- number of observations: $n = 500, 2000$
- responses are generated with overdispersion:

$$y_i \sim Po(s_i \exp(\eta_i)); s_i \sim U[0.66, 1.5]$$

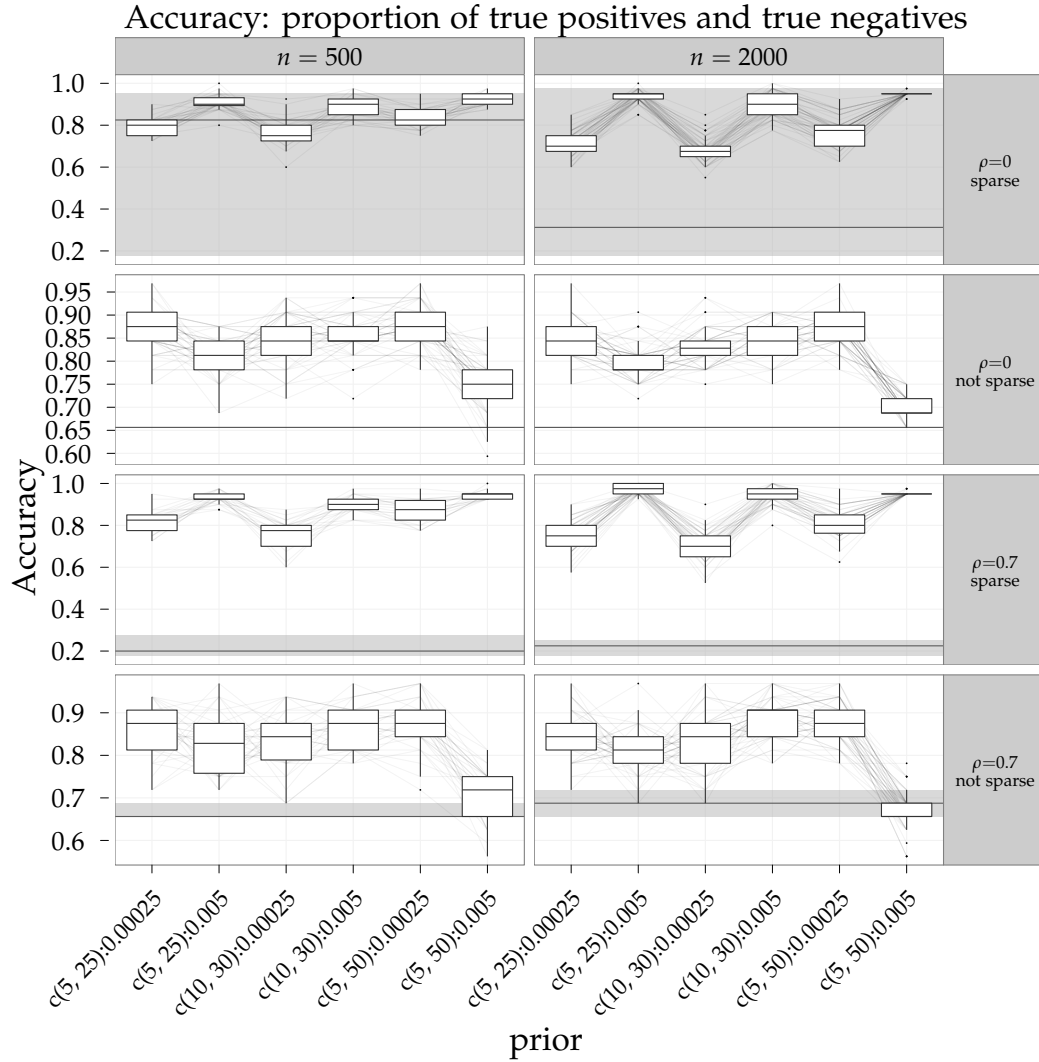


Figure 13: Complexity recovery for poisson response: proportion of correctly included and excluded model terms. Boxplots show results for the different prior settings. Horizontal ribbons show results for mboost: shaded region gives IQR, line represents median. Dark grey lines connect results for the same replication. Columns from left to right: 500 obs., 2000 obs. Rows from top to bottom: uncorrelated obs. with sparse and unsparse predictor, correlated obs. with sparse and unsparse predictor.

We did not use $v_0 = 0.01$ for this experiment because of its inferior performance in terms of complexity recovery in the Gaussian case.

Figure 12 shows the mean squared prediction error (on the scale of the linear predictor) divided by the one achieved by the “oracle”-GAM that includes only the relevant

covariates and no noise terms. Predictive performance is very robust against the different prior settings. Different prior settings also behave similarly within replications, as shown by the mostly parallel grey lines. Predictions are more precise than those of `ofmboost`, especially for smaller data sets (left column) and correlated responses (two bottom rows). For the “low sparsity, correlated covariates” setting (bottom row), the performance of our approach comes fairly close to that of the “oracle”-GAM, with relative prediction errors mostly between 1 and 1.5, and occasionally even improving on the oracle model for $n = 500$.

Figure 13 shows the proportion of correctly included and excluded terms (linear terms and basis expansions) in the estimated models. Estimated inclusion probabilities are sensitive to v_0 and comparatively robust against (a_τ, b_τ) . The smaller value for v_0 tends to perform better in the unsparsity settings (second and fourth rows) since it forces more terms into the model (resulting in higher sensitivity and lower specificity) and vice versa for the sparse setting and the larger v_0 . Complexity recovery is (much) better across the different settings and priors for our approach than for boosting. The constant accuracy for `mboost` in the low sparsity scenario with uncorrelated responses (second row) is due to its very low specificity: It includes practically all model terms all the time.

C.3. Gaussian GAM with concurvity

We use a similar data generating process as the one used in the previous Subsections:

- We define functions $f_1(x)$ to $f_4(x)$ as in the data generating process for the previous Subsections.
- We use 10 covariates: The first 4 are associated with functions f_1 to f_4 , respectively, while x_5 to x_{10} are “noise” variables without contribution to the linear predictor.
- covariates x are $\sim U[-2, 2]$
- we distinguish 3 scenarios of concurvity:
 - in scenario 1, $x_4 = c \cdot g(x_3) + (1 - c) \cdot u$, i.e., two covariates with real influence on the predictor are functionally related.

- in scenario 2, $x_5 = c \cdot g(x_4) + (1 - c) \cdot u$, i.e., a “noise” variable is a noisy version of a function of a covariate with direct influence.
- in scenario 3, $x_4 = c \cdot g(x_5) + (1 - c) \cdot u$, i.e., a covariate with direct influence is a noisy version of a “noise” variable.

where $g(x) = 2\Phi(x, \mu = -1, \sigma^2 = 0.16) + 2\Phi(x, \mu = 1, \sigma^2 = 0.09) - 4\phi(x) - 2$ with $\Phi(x, \mu, \sigma^2)$ defined as the cdf of the respective Gaussian distribution, i. i. d. standard normal variates u , and the parameter c controlling the amount of concurvity: $c = 1$ for perfectly deterministic relationship, and $c = 0$ for independence. In our simulation, $c = 0, .2, .4, .6, .8, 1$.

- we use signal-to-noise ratio SNR= 1, 5.
- we simulate 50 replications for each combination of the various settings.

Predictive MSE is evaluated on test data sets with 5000 observations. We use the default prior $a_\tau = 5, b_\tau = 25, v_0 = 0.00025$ for our approach and present results for 12 chains run in parallel for 2600 iterations each, discarding the first 100 as burn-in and keeping every fifth iteration. We compare predictive MSE and the sensitivity of the selection to various degrees of concurvity to results from

- mboost, as above,
- BART: Bayesian additive regression trees (Chipman et al., 2010) as implemented in R-package BayesTree (Chipman and McCulloch, 2010), as an example of a non-parametric method that does not yield interpretable models,
- hgam: the high-dimensional additive model approach of Meier et al. (2009), as implemented in R-package hgam
- spam: the approach for sparse additive models by Ravikumar et al. (2009), with a covariate assumed to be selected as a linear influence if its estimated associated degrees of freedom were between 0.1 and 1, and assumed to be selected as a non-linear influence if its estimated associated degrees of freedom were > 1 .

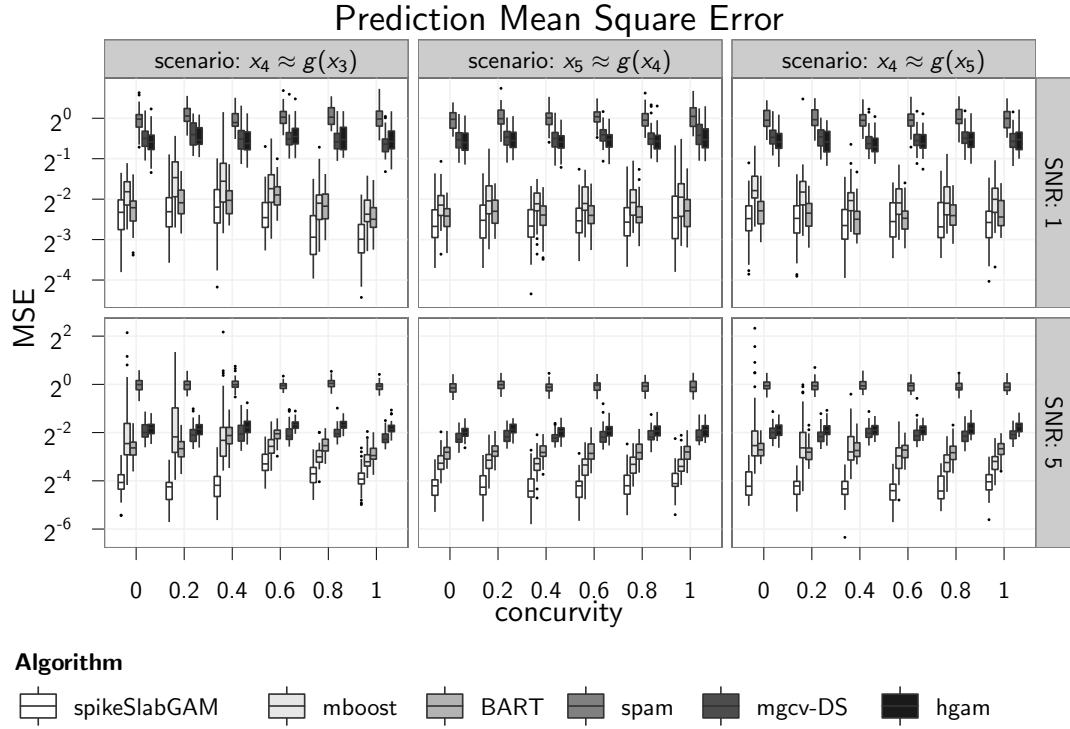


Figure 14: Prediction MSE. Boxplots show results for the different algorithms. Columns for the three scenarios, top row for signal to noise ratio 1, bottom row for signal to noise ratio 5. Vertical axis is on binary log scale.

- **mgcv-DS:** the double shrinkage approach for GAM estimation and term selection described in Marra and Wood (2011), as implemented in R-package *mgcv*, with a covariate assumed to be selected as a linear influence if its estimated associated degrees of freedom were between 0.1 and 1, and assumed to be selected as a non-linear influence if its estimated associated degrees of freedom were > 1 .

Figure 14 compares prediction MSEs for the various methods across scenarios and signal-to-noise ratios for varying degrees of concurvity. It is clear to see that our approach dominates in terms of prediction accuracy in these difficult settings, with BART and mboost as fairly close competitors. Figure 15 shows the proportion of correctly selected or removed covariates (i.e., “true positives” and “true negatives”). As for prediction accuracy, our approach outperforms the rest, with the double shrinkage approach of Marra

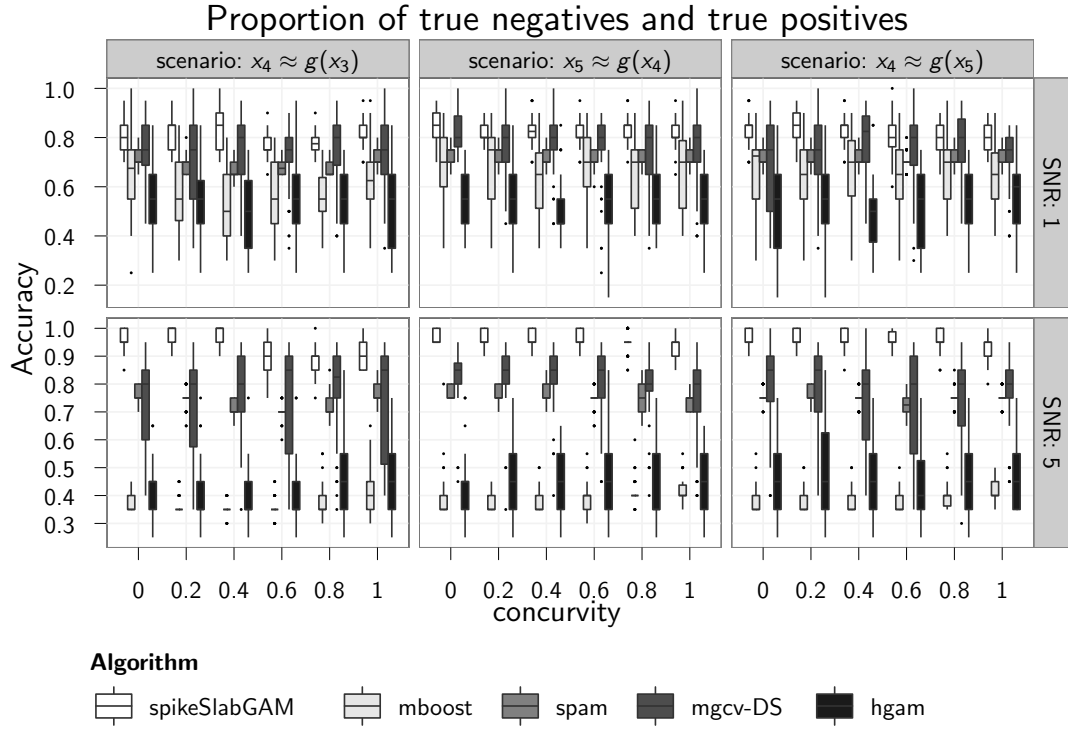


Figure 15: Selection accuracy as proportion of correctly selected or removed predictors. Boxplots show results for the different algorithms. Columns for the three scenarios, top row for signal to noise ratio 1, bottom row for signal to noise ratio 5.

and Wood (2011) a close second for selection accuracy for the noisy setting (but much worse in terms of prediction). Figure 15 does not include BART since its implementation in BayesTree does not offer clear inclusion or exclusion indicators. The only variable importance measure (i.e., how often a given variable was used in nodes across the ensemble of trees) returned by BART had roughly the same mean for influential and noise variables in all our simulations and would not have yielded a useable picture of the true predictor structure. Finally, figure 16 shows that estimated inclusion probabilities are fairly unreliable for intermediate to strong concurvity in noisy settings (top row) if both variables that are involved have separate effects (left column) – in this scenario, estimated inclusion probabilities of the covariate x_3 , which is a (noisy) function of another covariate x_4 , decrease dramatically for intermediate concurvity and recover somewhat for perfect

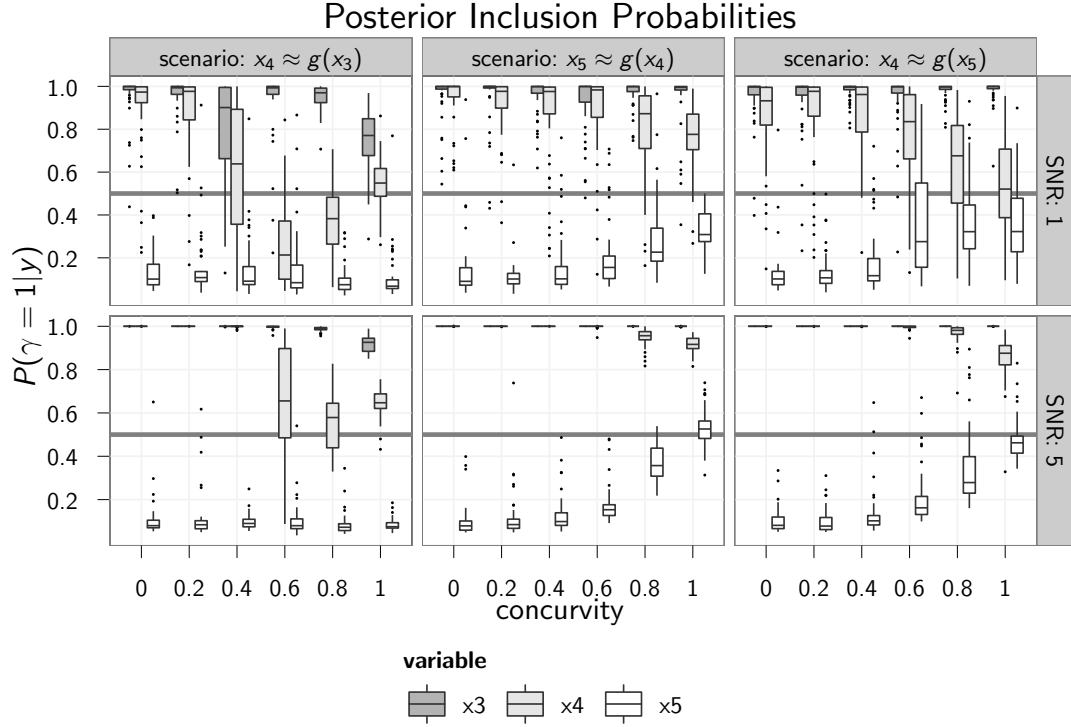


Figure 16: Posterior inclusion probabilities for x_3, x_4, x_5 . Shown are the respective maxima of the posterior inclusion probabilities for the linear and smooth effects of each variable in each replication. Columns for the three scenarios, top row for SNR 1, bottom row for SNR 5. Fat grey horizontal line denotes $P(\gamma = 1) = 0.5$

curvilinearity, where inclusion probabilities for x_3 are somewhat reduced. Note however, that if a model including interactions of x_3 and x_4 is specified in this scenario, our approach will often select those. This behavior makes sense if the effects of x_3 and x_4 cannot be clearly separated, as in this case.

If a variable x_5 without true effect is a (noisy) function of another covariate x_4 with true effect (second column), inclusion probabilities are fairly stable and yield correct inferences about the model structure unless the curvilinear relationship is perfect (concurvity= 1), at which point it again becomes impossible to disentangle the effect of x_4 and x_5 . Even so, using a threshold for inclusion of $P(\gamma = 1|y) > 0.5$ the correct model structure would have been identified in 44 out of 50 replicates for low SNR and 17 out of 50 for high SNR. For noisy data (top row), the scenario in which a noisy version x_4 of a spurious covariate

x_5 has true effects (third column) is problematic, with inclusion probabilities for x_4 decreasing strongly under intermediate and strong concavity. Note, however, that even for strong concavity ≥ 0.6 , the true model structure was identified at least 24 times out of 50 replicates for low SNR and at least 41 times out of 50 replicates for high SNR.

C.4. Summary

The simulations for generalized additive models show that the proposed peNMIG-Model is very competitive in terms of estimation accuracy and confirms that estimation results are robust against different hyperparameter configurations even in fairly complex models, and under strong concavity. Model selection is more sensitive towards hyperparameter configurations, especially v_0 . For smaller v_0 , spikeSlabGAM seems to be able to distinguish between important and irrelevant terms fairly reliably.

The performance of peNMIG as implemented in spikeSlabGAM seems to be very competitive to that of component-wise boosting as implemented in mboost and clearly dominates other function selection approaches in our concavity simulation study. Simulation results for an earlier, more rudimentary implementation of the peNMIG model on identical data generating processes and for many other settings are published in Scheipl (2010).

D. Additional Case Studies

According to German law, increases in rents for flats can be justified based on “average rents” paid for flats that are comparable in size, equipment, quality and location. As a consequence, most larger cities publish rental guides that provide such average rents, obtained from regression models with net rents or net rents per square meter as dependent variables.

Data: We analyze data on approximately 3,000 flats in Munich collected by Infratest Sozialforschung for the 2007 rental guide. The original data contain approximately 270 covariates describing characteristics of the flats as diverse as the quality of bathroom

equipment, whether the flat is rented for the first time, the presence and size of a garden or a balcony, etc. While most of the covariates are categorical, there are some continuous covariates of designated interest that are suspected to have nonlinear impact on the net rent per square meter. Moreover, a spatial variable is available that represents the subdistrict of Munich where the flat is located.

Model: We will model net rents per square meter with a high-dimensional geoadditive regression with Gaussian errors and linear predictor

$$\eta_i = \beta_0 + f_{\text{spat}}(s_i) + \sum_{j=1}^{267} f(x_{ij}),$$

where $f_{\text{spat}}(s_i)$ is the spatial effect of subdistrict s_i modeled with a Gaussian Markov random field (GMRF). The linear predictor additionally contains potential smooth effects of the year of construction, floorspace and begin of tenancy as well as 265 other potentially influential and mostly categorical covariates x_j . In total, this model has 594 coefficients in 269 model terms. Term selection is particularly challenging in this scenario because the available covariates include a number of redundant and highly collinear covariates such as an indicator variable for the presence of at least one balcony, a numeric variable giving the size of the flat's balconies in square meters and a count variable giving the number of balconies.

Hyperparameters were set to the default values determined in the simulation studies, i.e. $a_\tau = 5, b_\tau = 25, v_0 = 0.00025$ and $a_w = b_w = 1$. Estimates are based on 20 parallel chains running for 2500 iterations each after a burn-in of 500 iterations, with every fifth iteration saved.

Results: Table 3 lists the terms with posterior inclusion probability greater than 10%. The additive predictor is dominated by the contributions of terms for the presence of balcony, the date of beginning of the tenancy, the quality of the residential area the property is located in and presence of an attic and presence of a playground. Figures 17 and 18 show the estimated effect of subdistrict on net rent per square meter and the continuous predic-

Term	$P(\gamma = 1)$
MRF(subdistrict)	0.86
Floorspace, linear	0.95
Floorspace, smooth	0.97
Year, smooth	0.10
Begin of Tenancy, linear	1.00
Begin of Tenancy, smooth	0.97
Quality of Residential Area, cat.	0.58
Company Housing, cat.	0.82
Has Planted Area, cat.	0.26
Refrigerator, cat.	0.31
Kitchen Type, cat.	0.31
Intercom, cat.	0.80
Flooring, cat.	0.51
Has Balcony, cat.	0.48
Number of Balconies, cat.	0.21
Has Terrace, cat.	0.34
Number of Terraces, cat.	0.15
Has Roof Terrace, cat.	0.19
Number of Roof Terraces, cat.	0.12
Area of 2nd Balcony, linear	0.12
Has Clothes Drying Area, cat.	0.15
Has Playground, cat.	0.62
Has Attic, cat.	0.60
Garden-Use Permission, cat.	0.14
Apartment Type, cat.	0.15

Table 3: Terms with inclusion probability $P(\gamma = 1) > 0.1$.

tor effects, respectively, both combined with associated credible intervals. In general, the estimated effects resemble those found in previously published analyses (cf. Kneib et al., 2011) of this data, with lower rents in predominantly working-class outskirts and higher rents in fashionable districts in and around the city centre. Including the beginning of tenancy (which has not been included in the analyses by Kneib et al., 2011) seems to somewhat attenuate the effect of the year of construction of the flat, and the total floorspace has a much larger effect on net rent per square meter.

Cross Validation Performance: We perform a 10-fold crossvalidation to gain some insight into the stability of the term selection and to compare the predictive performance of our approach with previous modeling efforts described in Kneib et al. (2011). While ridge and lasso priors to select single dummy variables of specific factor levels have been used in their model, our results are based on blockwise selection including or excluding all levels of a categorical covariate simultaneously. Models in Kneib et al. (2011) were estimated

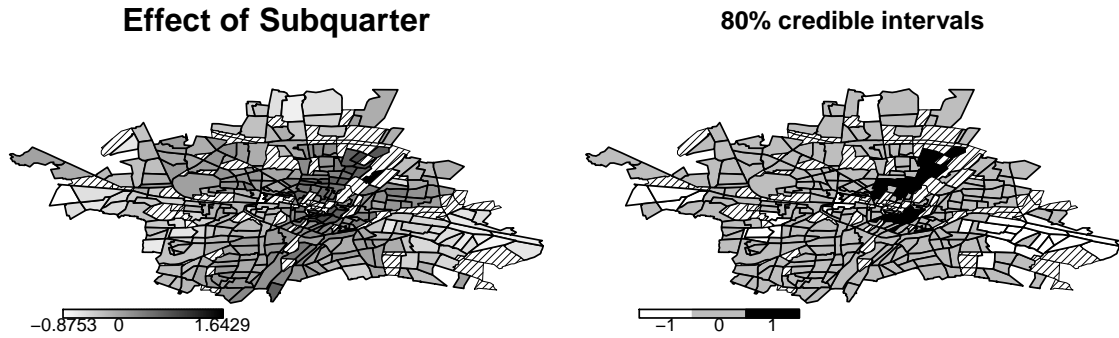


Figure 17: Map of Munich's subdistricts with estimated effect on net rent/m². Left panel shows posterior mean of effects, right panel shows the sign of 80% posterior credible intervals: regions with lower net rent in white, higher net rent in black, regions with credible intervals overlapping zero in gray. Subdistricts without any observations are filled with diagonal lines.

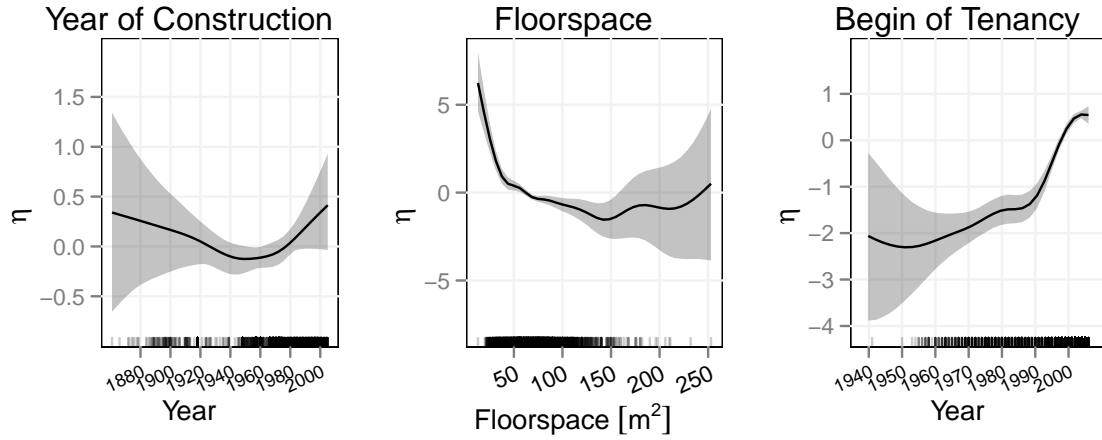


Figure 18: Smooth terms with posterior inclusion probability greater than 10% for Munich rental guide data. Grey ribbons show 80% pointwise credible intervals.

with the software package BayesX (Brezger et al., 2005). We also considered “expert” models in analogy to Kneib et al. (2011) including only a strongly reduced number of covariates as candidates, estimated both with BayesX and our function selection approach. For the latter, we also estimated an expanded “expert” model including all potential two-way interactions (except those with the spatial GMRF term). The resulting model has 1051

coefficients in 190 model terms including varying coefficient terms and smooth interaction surfaces.

Figure 19 displays the cross validation error on the ten folds for the different model specifications. We see that prediction accuracy is not diminished by putting a model

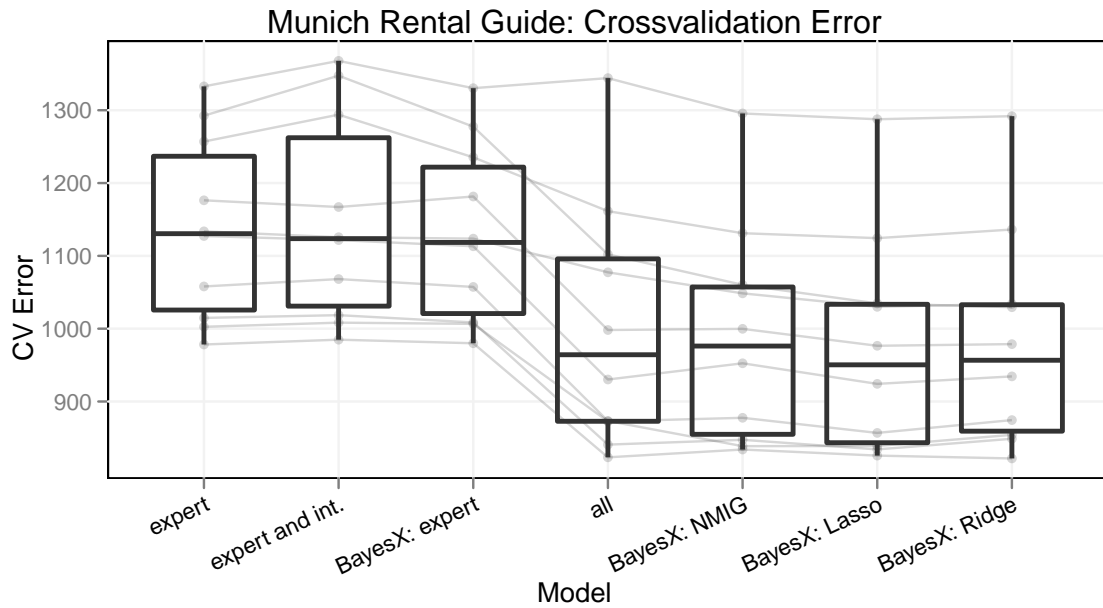


Figure 19: Prediction error for 10 cross validation folds for the Munich rental guide data. Grey lines connect results from identical folds.

with only relevant terms under selection, as shown by the equivalent performances of the peNMIG “expert”-model and that of the “expert”-model fit with BayesX (first and third boxes from the left). This indicates that the relevant effects are estimated without bias despite the variable selection prior associated with them, a consequence of the adaptive shrinkage properties of the peNMIG prior. As for the sepsis survival data analyzed in the subsequent section, there is no noticeable change in prediction performance in most folds if a large number of interaction terms are added — compare the similar performances of the expert model and of the expert model with additional two-way interactions (first and second boxes from the left). The underlying reason is that only a single interaction effect, the interaction between the level of kitchen furnishing and level of pollution has $P(\gamma = 1) > 0.1$ in the expanded “expert” model, but its effect is still very small. The

precision of the prediction achieved by our peNMIG approach on the full data set with 269 potential model terms (fourth box from the left) is very similar to that of Bayesian lasso, Bayesian ridge and the conventional NMIG approach (as implemented in BayesX) for most folds, even though estimates of the dominating (nonlinear) effects for floorspace, year of construction and begin of tenancy and the effect of subquarter were associated with a variable selection prior in our model while they were associated with conventional smoothing priors in BayesX. This reinforces our conclusion that important effects are estimated without a selection-induced attenuating bias in our approach.

Variable selection is very stable across the folds, with the same terms included in all ten folds for the “expert” model and the “expert” model with interactions. Only a single one of the interactions (between kitchen furnishings and level of pollution), has marginal posterior inclusion probabilities above 0.1 in six of the ten folds, the rest is excluded unequivocally in all folds. The model with interactions is more conservative and includes less of the main effect terms than the smaller model, because the large number of irrelevant terms moves most of the posterior mass for w , the overall prior inclusion probability, towards very low values: the posterior means of w in the smaller model are between 0.71 and 0.85, while the posterior means for w in the model with interactions are between 0.07 and 0.09. In the model with all possible covariates, a core set of 23 covariates is identified in at least nine out of the ten folds, while nine other covariates have marginal posterior inclusion probabilities above 0.1 in at least one fold. Of those nine, two do so in eight of the folds.

D.1. Case Study: Hymenoptera Venom Allergy

Data We reanalyze data on bee and wasp venom allergy from a large observational multicenter study previously analyzed in Rüeff et al. (2009). The data consists of 962 patients from 14 European study centers with established bee or wasp venom allergy who suffered an allergic reaction after being stung. The binary outcome of interest is whether patients suffered a severe, life-threatening reaction, defined as anaphylactic shock, loss of consciousness, or cardiopulmonary arrest. A severe reaction was observed for 206 of

the 962 patients (21.4%). Data were collected on the concentration of tryptase, a potential biomarker, patients' sex and age, whether the culprit insect was a bee or wasp, on the intake of three types of cardiovascular medication (β -blockers, ACE inhibitors and anti-hypertensive drugs), whether the patient had had at least one minor systemic reaction to a sting prior to the index sting and the CAP-class (a measure of antibody load) of the patient with regard to the venom of the culprit insect, with levels 0, 1, 2, 4, 5+.

Models An analysis of this data has to take into account possible study center effects, possible non-linear effects of both age and the (logarithm of) blood serum tryptase concentrations and the possibility of differing effect structures for bee and wasp stings. Our aim is twofold again: We want to (1) estimate a model that allows assessment of the influence of each covariate on the susceptibility for a severe reaction, accounting for possibly nonlinear effects and interaction effects and (2) use this setting to evaluate the stability of the selection and estimation of increasingly complex models on real data as well as investigate the consequences of less-than-optimal sampler convergence we observed.

Full Data Analysis We fit a peNMIG model with all main effects and all second order interactions except those with study centre, with smooth functions for both age and tryptase and a random intercept for the study center. In total, this model has 267 coefficients in 66 model terms: 13 main effects including the global intercept, separate linear and non-linear terms for age and tryptase and a random intercept for study centre, 21 interactions between the seven factor variables, 28 terms for the linear and smooth interactions for age and tryptase with each of the seven factors, and four terms for the interaction effect of age and tryptase (one linear-linear interaction, two varying coefficient terms, one smooth interaction surface) Results are based on samples from 20 chains with 40000 iterations each after 1000 burn-in, with every 20th saved. Running a single chain of this length on a modern desktop computer (i.e., Intel Q9550 2.83GHz) takes about 45 minutes, so that the entire fit takes about 4 hours on a quad-core CPU.

Figure 20 shows the estimated effects of the terms with $P(\gamma = 1) > .1$ that are listed in Table 4. Since the inclusion probabilities indicate interlocking interactions of cap, tryptase

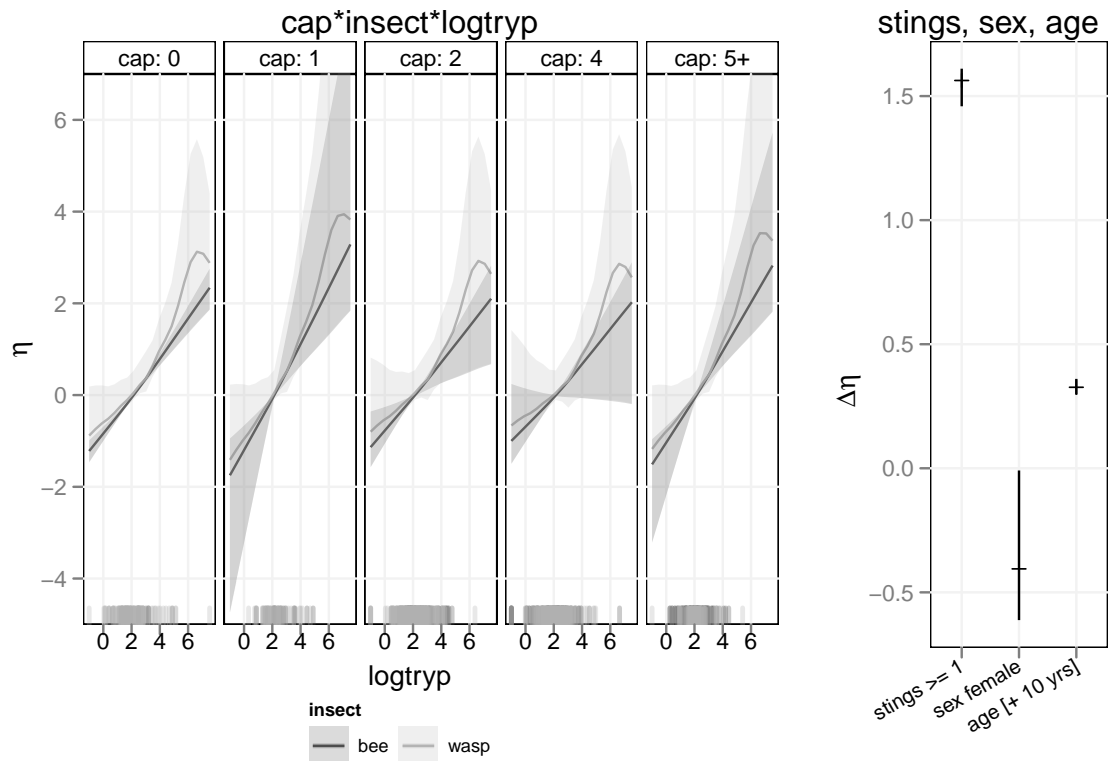


Figure 20: Posterior means of effects with (pointwise) 80% credible intervals. Only effects for terms with marginal inclusion probability $> .1$ are shown. Since there is some evidence for interlocking interactions of cap class, tryptase and culprit insect (c.f. Table 4), the left graph shows the joint effect of these 3 variables. The graph on the right shows the relative effects of previous stings (compared to none before the index sting), female gender (compared to male) and an increase in the patient's age by 10 years.

and culprit insect, the panels in the left graph in the figure show the joint effects of these 3 variables. Each panel shows the effect estimate of tryptase plasma concentration for bee patients (dark grey) and wasp patients (light grey) for the given CAP class. The rug plot at the bottom indicates the locations of the data. The large uncertainty precludes a detailed interpretation of this 3-way interaction, but in general, the risk is higher for wasp patients: the main effect of culprit insect yields an odds ratio of 1.16 (80%CI: 1-2.43) and the increase in risk in wasp patients seems to be smaller for lower and larger for higher tryptase concentrations. The graph on the right shows the relative effects of previous stings (compared to none before the index sting), female gender (compared to male) and an increase in the patient's age by 10 years. Estimated random effects for the

Term	$P(\gamma = 1)$
culprit insect	0.16
stings	1.00
sex	0.70
age, linear	1.00
tryptase, log-linear	1.00
study centre	0.71
insect:tryptase, smooth	0.46
cap:tryptase, log-linear	0.14

Table 4: Posterior means of marginal inclusion probabilities $P(\gamma = 1)$ (only given for terms with $P(\gamma = 1) > .1$).

study centres are not shown, their associated posterior mean odds ratios range between 0.44 and 2.13.

Lack of Convergence for γ For this fairly complicated model, we experience some difficulties with the convergence of the MCMC sampler: We observe poor mixing for some of the entries in γ , with chains getting stuck in basins of attraction around posterior modes for long periods of time. This leads to posterior inclusion probabilities for single chains often ending up either close to zero or close to one for some of the terms. Running a large number of parallel chains from random starting configurations seems to remedy this problem. To investigate this issue, we perform a large MCMC experiment with 800 chains, each with 10000 iterations after 100 burnin, for the model described above. Figure 21 shows the average inclusion probabilities for the 16 terms with the highest between-chain variability of $P(\gamma = 1)$ for 20 fits with 40 chains each. Grey lines connect posterior means based on an increasing number of chains for each fit. The black horizontal line shows the mean over all 800 chains, which we presume to be a good estimate of the “true” marginal posterior inclusion probability. Convergence of the posterior means is slow for these terms, but discrimination between important, intermediate and negligible effects seems to be reliable based on as few as 10 to 20 chains. While we would not be comfortable in claiming that 10 or 20 parallel chains are enough to completely explore this very high-dimensional model space and yield a reliable estimate of posterior model probabilities, i.e., the joint distribution of γ , the marginal inclusion probabilities $P(\gamma_j = 1)$, $j = 1, \dots, p$

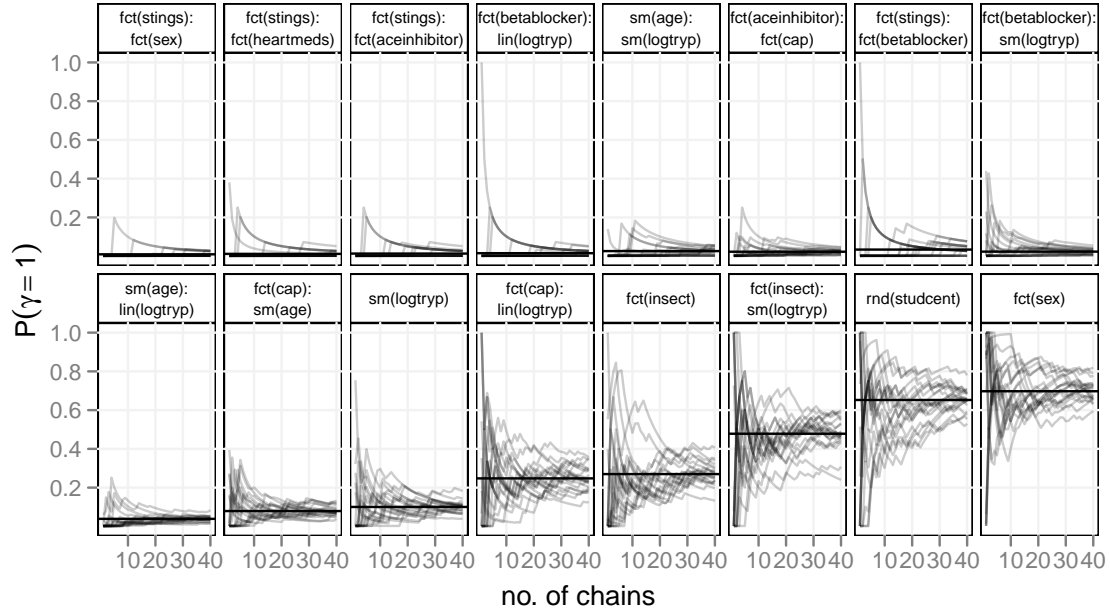


Figure 21: Average inclusion probabilities for those terms with convergence issues for 20 fits with 40 chains each. Grey lines connect posterior means over an increasing number of chains for each fit. Black horizontal line shows the mean over all 800 chains.

of the various terms seem to be estimated well enough to distinguish between important, intermediate and negligible effects, which is usually all that is required in practice. This conclusion is also borne out by the UCI benchmark study in the main article.

Predictive Performance Comparison We subsample the data 20 times to construct independent training data sets with 866 subjects each and test data sets with the remaining 96 patients to evaluate the precision of the resulting predictions and compare predictive performance to that of equivalent component-wise boosting models fitted with `mboost` and an unregularized GAMM-fit with all main effects estimated with `gamm4`. Results for our approach are based on 8 parallel chains each running for 10000 iterations after 500 iterations of burn-in, with every 10^{th} iteration saved. Component-wise boosting results are based on a stopping parameter determined by a 25-fold bootstrap of the training data, with a maximal iteration number of 500. We compare three model specification of increasing complexity: a simple model with main effects only, a model with main effects and all interactions between culprit insect and the other covariates, and the complex model

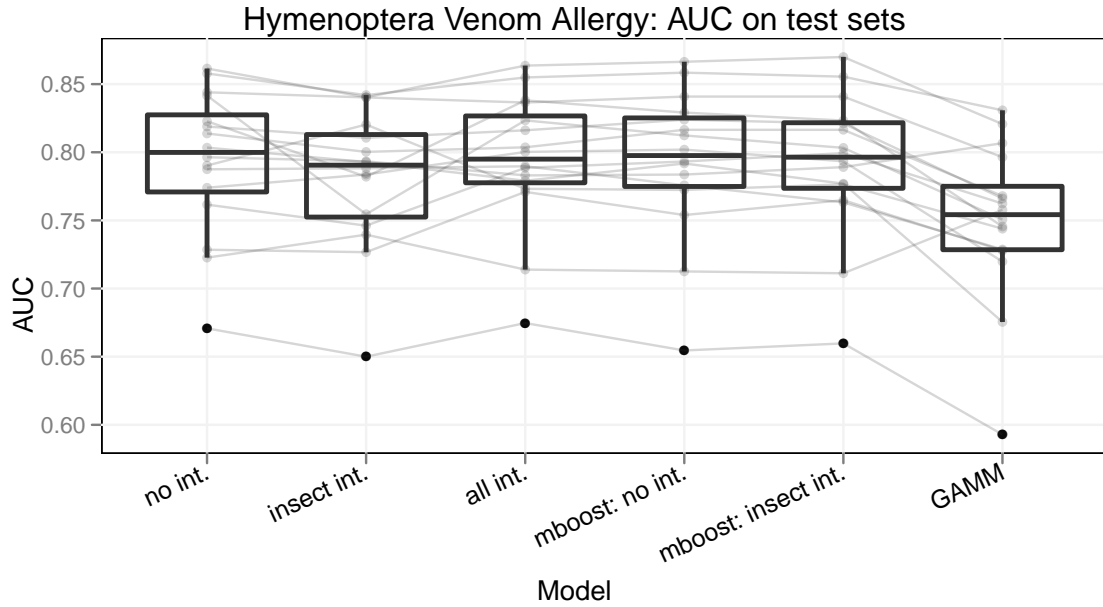


Figure 22: Area under the ROC curve for 20 test sets from the hymenoptera venom allergy data set, higher is better. Grey lines connect results from identical folds.

with all main effects and second order interactions presented before. We were unable to fit the latter model with mboost, and the model including the insect interactions could not be fitted by mboost for 4 of the training data sets. We report results for the 16 sets remaining. Figure 22 shows the area under the ROC curve (AUC) achieved by the different model specifications. For this data set, the models with higher maximal complexity show slight decreases in predictive accuracy, but still perform better than an unregularized generalized additive mixed model (GAMM) on the far right.

Despite the fairly low number of parallel chains and comparatively short chain lengths, the stability of the marginal term inclusion probabilities across subsamples is fairly good, indicating that the term selection is robust to small changes in the data and that even as few as 8 chains may be enough to reach fairly reliable rough estimates of term importance in this difficult setting. All model specifications identified the same subset of important main effects (i.e., number of previous stings before the index sting, sex, linear effects of age and the log of tryptase and the random effect for study centre). Figure 23 shows the posterior means of inclusion probabilities $P(\gamma = 1)$ across 16 subsampled training data

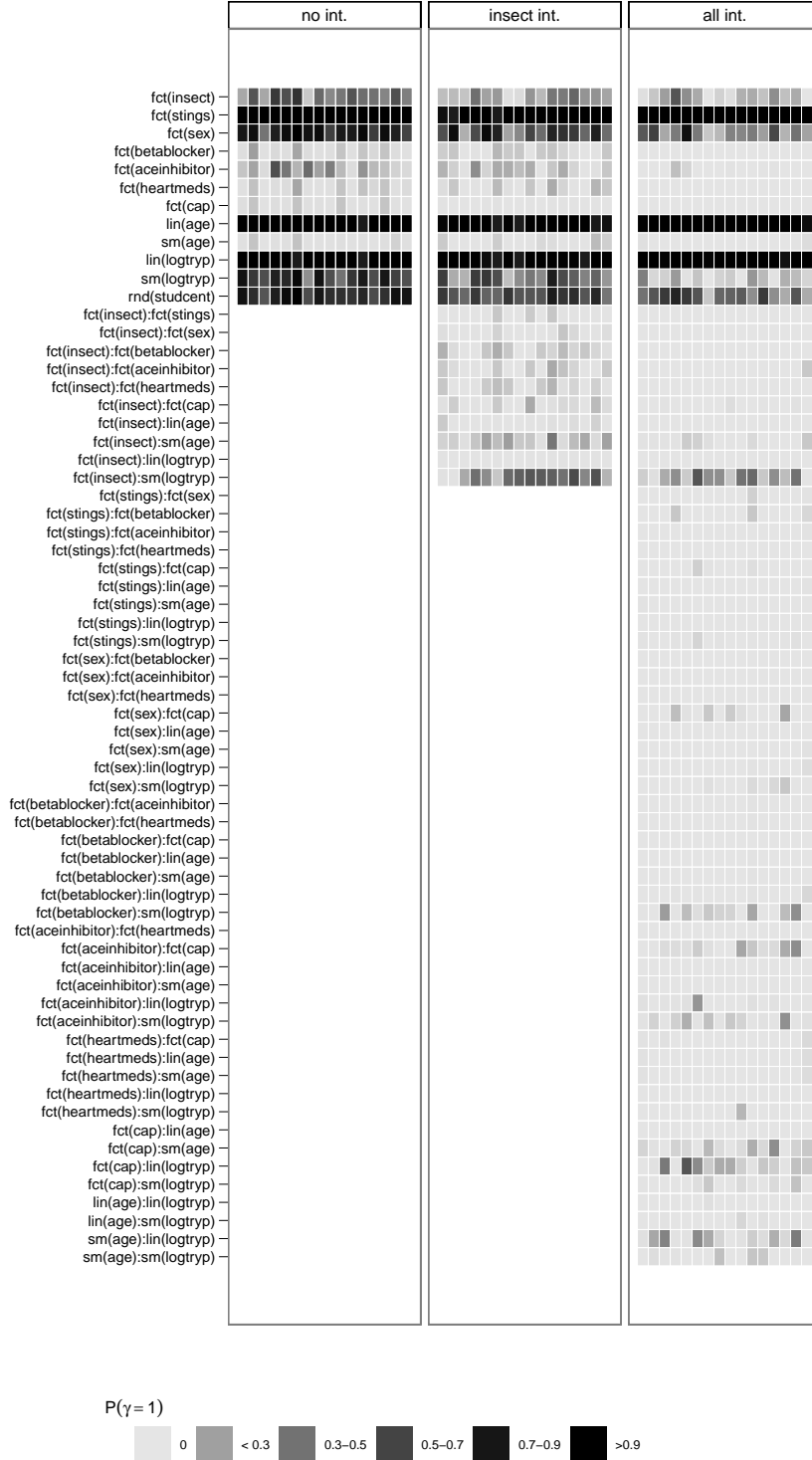


Figure 23: Posterior means of inclusion probabilities $P(\gamma = 1)$ across 16 subsampled training data sets for the 3 model specifications.

sets for each of the 3 model specifications.

References

- Bae, K. and B. Mallick (2004). Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics* 20(18), 3423–3430.
- Baglama, J. and L. Reichel (2006). Augmented implicitly restarted lanczos bidiagonalization methods. *SIAM Journal on Scientific Computing* 27(1), 19–42.
- Brezger, A., T. Kneib, and S. Lang (2005). BayesX: Analyzing Bayesian structural additive regression models. *Journal of Statistical Software* 14(11).
- Brezger, A. and S. Lang (2006). Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics & Data Analysis* 50(4), 967–991.
- Bühlmann, P. and T. Hothorn (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science* 22, 477–505.
- Carvalho, C., N. Polson, and J. Scott (2010). The horseshoe estimator for sparse signals. *Biometrika* 97(2), 465–480.
- Chipman, H., E. George, and R. McCulloch (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics* 4(1), 266–298.
- Chipman, H. and R. McCulloch (2010). *BayesTree: Bayesian Methods for Tree Based Models*. R package version 0.3-1.1.
- Cottet, R., R. Kohn, and D. Nott (2008). Variable Selection and Model Averaging in Semiparametric Overdispersed Generalized Linear Models. *Journal of the American Statistical Association* 103(482), 661–671.
- Crainiceanu, C., D. Ruppert, G. Claeskens, and M. P. Wand (2005). Exact likelihood ratio tests for penalised splines. *Biometrika* 92, 91–103.
- Crainiceanu, C. M., D. Ruppert, and M. P. Wand (2005). Bayesian analysis for penalized spline regression using WinBUGS. *Journal of Statistical Software* 14.
- Denison, D., B. Mallick, and A. Smith (1998). Bayesian MARS. *Statistics and Computing* 8(4), 337–346.
- Dimatteo, I., C. R. Genovese, and R. E. Kass (2001). Bayesian curve-fitting with free-knot splines. *Biometrika* 88, 1055–1071.

- Fahrmeir, L., T. Kneib, and S. Lang (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica* 14, 731–761.
- Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer.
- Gelman, A., D. Van Dyk, Z. Huang, and J. Boscardin (2008). Using redundant parameterizations to fit hierarchical models. *Journal of Computational and Graphical Statistics* 17(1), 95–122.
- George, E. and R. McCulloch (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* 88(423), 881–889.
- Greven, S., C. Crainiceanu, H. Küchenhoff, and A. Peters (2008). Restricted likelihood ratio testing for zero variance components in linear mixed models. *Journal of Computational and Graphical Statistics* 17(4), 870–891.
- Hofner, B., T. Kneib, W. Hartl, and H. Küchenhoff (2011). Building cox-type structured hazard regression models with time-varying effects. *Statistical Modelling* 11, 3–24.
- Hothorn, T., P. Buehlmann, T. Kneib, M. Schmid, and B. Hofner (2010). *mboost: Model-Based Boosting*. R package version 2.0-0.
- Huang, J., J. Horowitz, and F. Wei (2010). Variable selection in nonparametric additive models. *Ann. Statist.* 38, 2282–2313.
- Ishwaran, H. and J. Rao (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics* 33(2), 730–773.
- Kneib, T., S. Konrath, and L. Fahrmeir (2011). High-dimensional structured additive regression models: Bayesian regularization, smoothing and predictive performance. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 60, 51–70.
- Krivobokova, T., C. M. Crainiceanu, and G. Kauermann (2008). Fast adaptive penalized splines. *Journal of Computational and Graphical Statistics* 17, 1–20.
- Lang, S. and A. Brezger (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics* 13(1), 183–212.
- Lewis, B. (2009). *irlba: Fast Partial SVD by Implicitly-Restarted Lanczos Bidiagonalization*. R package version 0.1.1.

- Lin, Y. and H. Zhang (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.* 34, 2272–2297.
- Marra, G. and S. Wood (2011). Practical variable selection for generalized additive models. *Computational Statistics and Data Analysis* 55, 2372–2387.
- Meier, L., S. van der Geer, and P. Bühlmann (2009). High-dimensional additive modeling. *Ann. Statist.* 37, 3779–3821.
- Meyer, D., F. Leisch, and K. Hornik (2003). The support vector machine under test. *Neurocomputing* 55(1-2), 169–186.
- Mitchell, T. and J. Beauchamp (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association* 83(404), 1023–1032.
- Morris, J. S. and R. J. Carroll (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series B* 68(2), 179–199.
- O’Hara, R. and M. Sillanpää (2009). A Review of Bayesian Variable Selection Methods: What, How, and Which? *Bayesian Analysis* 4(1), 85–118.
- Panagiotelis, A. and M. Smith (2008). Bayesian identification, selection and estimation of semiparametric functions in high-dimensional additive models. *Journal of Econometrics* 143(2), 291–316.
- Polson, N. and J. Scott (2010). Shrink globally, act locally: Sparse bayesian regularization and prediction. In J. Bernardo, M. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West (Eds.), *Bayesian Statistics 9*. Oxford University Press.
- Ravikumar, P., H. Liu, J. Lafferty, and L. Wasserman (2009). Sparse additive models. *J. Roy. Statist. Soc. Ser. B* 71, 1009–1030.
- Reich, B., C. Storlie, and H. Bondell (2009). Variable selection in Bayesian smoothing spline ANOVA models: Application to deterministic computer codes. *Technometrics* 51(2), 110.
- Rue, H. and L. Held (2005). *Gaussian Markov Random Fields*. Chapman & Hall / CRC.
- Ruëff, F., B. Przybilla, M. Biló, U. Müller, F. Scheipl, W. Aberer, J. Birnbaum, A. Bodzenta-Lukaszyk, F. Bonifazi, C. Bucher, et al. (2009). Predictors of severe systemic anaphylactic reactions in patients with Hymenoptera venom allergy: Importance of baseline serum tryptase—a study of the European Academy of Allergology and Clinical Immunology Interest Group on Insect Venom Hypersensitivity. *Journal of Allergy and Clinical Immunology* 124(5), 1047–1054.

- Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression*. Cambridge University Press.
- Sabanés Bové, D. and L. Held (2011). Hyper-g Priors for Generalized Linear Models. *Bayesian Analysis*. to appear.
- Sabanés Bové, D., L. Held, and G. Kauermann (2011). Hyper-g Priors for Generalised Additive Model Selection with Penalised Splines. submitted.
- Scheipl, F. (2010). Normal-mixture-of-inverse-gamma priors for bayesian regularization and model selection in generalized additive models. Technical Report 84, Department of Statistics, LMU München.
- Scheipl, F. (2011a). *Bayesian Regularization and Model Choice in Structured Additive Regression*. Dr. Hut Verlag.
- Scheipl, F. (2011b, 9). spikeSlabGAM: Bayesian variable selection, model choice and regularization for generalized additive mixed models in R. *Journal of Statistical Software* 43(14), 1–24.
- Scheipl, F., S. Greven, and H. Küchenhoff (2008). Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics & Data Analysis* 52(7), 3283–3299.
- Storlie, C., H. Bondell, B. Reich, and H. Zhang (2010). Surface estimation, variable selection, and the nonparametric oracle property. *Statistica Sinica*. in press.
- Wand, M. P. (2000). A comparison of regression spline smoothing procedures. *Computational Statistics* 15, 443–462.
- Wood, S. (2008). Fast stable direct fitting and smoothness selection for generalized additive models. *Journal of the Royal Statistical Society. Series B* 70(3), 495.
- Wood, S., R. Kohn, T. Shively, and W. Jiang (2002). Model selection in spline nonparametric regression. *Journal of the Royal Statistical Society. Series B* 64(1), 119–139.
- Yau, P., R. Kohn, and S. Wood (2003). Bayesian variable selection and model averaging in high-dimensional multinomial nonparametric regression. *Journal of Computational and Graphical Statistics* 12(1), 23–54.

Zhu, H., P. Brown, and J. S. Morris (2011). Robust, adaptive functional regression in functional mixed model framework. *Journal of the American Statistical Association*. to appear.

Zhu, H., M. Vannucci, and D. Cox (2010). A bayesian hierarchical model for classification with selection of functional predictors. *Biometrics* 66(2), 463–473.